



# UCAM

UNIVERSIDAD CATÓLICA  
DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO  
Programa de Doctorado Tecnologías de las Computación e  
Ingeniería Ambiental

Desarrollo de herramientas bioinformáticas fácilmente  
usables y accesibles vía web con aplicabilidad general en  
contextos farmacológicos, agrícolas, nutracéuticos y  
cosméticos.

**Autor:**

Antonio Jesús Banegas Luna

**Directores:**

Dr. D. Horacio Emilio Pérez Sánchez

Dr. D. José Pedro Cerón Carrasco

Murcia, junio de 2019





# UCAM

UNIVERSIDAD CATÓLICA  
DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO  
Programa de Doctorado Tecnologías de la Computación e  
Ingeniería Ambiental

Desarrollo de herramientas bioinformáticas fácilmente  
usables y accesibles vía web con aplicabilidad general en  
contextos farmacológicos, agrícolas, nutracéuticos y  
cosméticos.

**Autor:**

Antonio Jesús Banegas Luna

**Directores:**

Dr. D. Horacio Emilio Pérez Sánchez

Dr. D. José Pedro Cerón Carrasco

Murcia, junio de 2019





# UCAM

UNIVERSIDAD CATÓLICA  
DE MURCIA

## AUTORIZACIÓN DE LO/S DIRECTOR/ES DE LA TESIS PARA SU PRESENTACIÓN

El Dr. D. Horacio Emilio Pérez Sánchez y el Dr. D. José Pedro Cerón Carrasco como Directores de la Tesis Doctoral titulada “Desarrollo de herramientas bioinformáticas fácilmente usables y accesibles vía web con aplicabilidad general en contextos farmacológicos, agrícolas, nutracéuticos y cosméticos” realizada por D. Antonio Jesús Banegas Luna en el Departamento de Tecnología de la Computación e Ingeniería Ambiental, **autorizan su presentación a trámite** dado que reúne las condiciones necesarias para su defensa.

Lo que firmamos, para dar cumplimiento al Real Decreto 99/2011, 1393/2007, 56/2005 y 778/98, en Murcia a 6 de junio de 2019.

Dr. D. Horacio Emilio Pérez  
Sánchez

Dr. D. José Pedro Cerón Carrasco

UCAM



**EIDUCAM**  
Escuela Internacional  
de Doctorado



## RESUMEN

El cribado virtual (*virtual screening*, VS) es una técnica computacional, empleada frecuentemente en bioinformática, cuyo objetivo es reducir un espacio químico de gran tamaño y complejidad en otro más reducido y manejable. Para esta tarea, el cribado virtual basado en ligandos (*ligand-based virtual screening*, LBVS) se ha convertido en una alternativa eficiente frente a otras técnicas más complejas computacionalmente tales como la Dinámica Molecular, puesto que permite encontrar rápidamente los compuestos más prometedores a un bajo coste computacional. En los últimos años, con el desarrollo del big data y de los paradigmas de supercomputación, han emergido numerosos servidores web que prestan servicios de LBVS y que aprovechan la computación de alto rendimiento (*high performance computing*, HPC) para mejorar sus prestaciones. Pero, a pesar de que la mejora del rendimiento es un punto importante, el principal factor de calidad de estos servidores sigue siendo la fiabilidad de sus predicciones.

Esta tesis pretende analizar las características de los servidores de LBVS actuales, prestando especial atención a las tecnologías que emplean y al rendimiento que, en términos computacionales, extraen de las plataformas HPC. Una vez obtenidas las conclusiones de dicho análisis, el objetivo es desarrollar una herramienta web que solvete las debilidades de los servidores existentes aplicando técnicas no estudiadas hasta ahora en el campo de LBVS.

Como resultado, se presentan los estudios teóricos realizados y la herramienta web BRUSELAS (*Balanced, Rapid and Unrestricted Server for Extensive Ligand-Aimed Screening*), la cual está disponible, de manera totalmente gratuita, en <http://bio-hpc.eu/software/Bruselas>. Como principales características diferenciadoras de BRUSELAS destacan la utilización de funciones de consenso para combinar las predicciones de varios algoritmos de similitud y farmacofóricos, el uso de descriptores moleculares y palabras clave para crear librerías dinámicamente y el filtrado de los resultados mediante filtros moleculares. Además, se ha implementado un potente módulo de análisis de resultados que permite su procesamiento tanto *online* como *offline*, mediante la conocida herramienta PyMOL, y la visualización de la salida generada por cada

algoritmo de similitud. La nueva herramienta se ha aplicado a casos de estudio prácticos, como la búsqueda de anticoagulantes sanguíneos y de posibles fármacos para terapias contra el cáncer. Las tareas ejecutadas han dado lugar a interesantes resultados teóricos que pueden servir como base para la experimentación en etapas posteriores, ya sea *in vitro* o *in vivo*.

En conclusión, se puede afirmar que BRUSELAS puede ser una herramienta muy útil en la etapa de búsqueda de compuestos candidatos a fármacos. Además, BRUSELAS proporciona a los usuarios nuevas funcionalidades que otros servidores web no proporcionan a través de una interfaz amigable y sin necesidad de tener grandes conocimientos en informática. Los resultados acumulados confirman que es una arquitectura fiable en cuanto a la calidad de las predicciones, y que tiene un rendimiento comparable al de otros servidores similares. Por lo tanto, BRUSELAS puede ser de gran ayuda en futuros estudios, ya sea utilizada de manera individual o colaborando con otras técnicas computacionalmente más costosas (p.ej. *docking*).

**Palabras clave:** Descubrimiento de fármacos, Cribado virtual, Química computacional, Servidor web, Computación de alto rendimiento, Similitud molecular, Cribado farmacofórico.



## ABSTRACT

Virtual screening (VS) is a computational technique, frequently used in bioinformatics, whose main goal is to reduce a large and complex chemical space into a smaller and less complex one. To achieve so, the ligand-based virtual screening (LBVS) approach has become a cost-effective alternative against other more computationally demanding approaches such as Molecular Dynamics, since it allows us to quickly find the most promising compounds with a low computational cost. In recent years, with the development of the big data and the supercomputing paradigms, several web servers providing LBVS services have emerged. These servers benefit from high performance computing (HPC) to improve their performance. However, although performance improvement is an important point, the crucial factor in evaluating the quality of these servers remains the reliability of their predictions.

This thesis aims to analyse the skills of current LBVS servers, paying special attention to the underlying technologies and the performance they obtain from HPC platforms. Once the main conclusions are obtained, the objective is to develop a web tool to overcome the weaknesses of the existing servers by applying techniques still unexplored in the field of LBVS.

As a result, the theoretical studies and the web tool BRUSELAS (*Balanced, Rapid and Unrestricted Server for Extensive Ligand-Aimed Screening*) are presented. BRUSELAS is available free of any cost at <http://bio-hpc.eu/software/Bruselas>. The most distinctive features of BRUSELAS are the use of consensus scoring functions to combine the scores of many similarity and pharmacophoric algorithms, the use of molecular descriptors and keywords to dynamically create chemical libraries and the filtering of results using molecular filters. In addition, a powerful result analysis module has been implemented, which allows processing the output both online and offline, by means of the well-known PyMOL tool, and the visualization of the alignments generated by each similarity algorithm. The new tool has been used in case studies, such as the search for blood anticoagulants and the search for potential drugs to be applied in cancer therapies. The tasks that were carried out resulted in interesting theoretical results

that can lead to a more detailed research in the following stages, either in vivo or in vitro.

In summary, it can be said that BRUSELAS can be a very useful tool in the step of searching for drug candidates. In addition, BRUSELAS provides users with novel functionalities not yet provided by other web servers through a friendly web interface and without the need to be an expert on informatics. The results confirm that the new architecture is reliable enough in terms of the accuracy of the predictions, and outperforms some of its competitors. Therefore, BRUSELAS can help in future research, either as an independent method or collaborating with other approaches (e.g. docking).

**Keywords:** Drug discovery, Virtual screening, Computational chemistry, Web server, High performance computing, Similarity searching, Pharmacophore screening.

## AGRADECIMIENTOS

No sería justo que no comenzara esta sección agradeciendo toda su ayuda a las dos personas más importantes que hay detrás de esta tesis: Horacio y José Pedro. Horacio confió en mí incluso antes de conocernos personalmente. Un par de correos y unas llamadas fueron suficientes para que se aventurase a dirigir una tesis a casi 2.000 kilómetros de distancia. Desde entonces lo he conocido un poco más y sólo puedo decir cosas buenas de él. Siempre ha estado a mi lado en los momentos más duros, aportando ideas y guiando la investigación por el mejor camino. Y si hoy estoy aquí es en buena parte gracias a él. No menos agradecido estoy a José Pedro. No dudó en unirse al equipo cuando apenas estábamos empezando y ha sido una pieza fundamental. Además de ser dos magníficos investigadores, son dos grandes personas. Y para no extenderme mucho, sólo voy a decir que no podría haber deseado unos directores mejores. También quiero hacer extensivo ese agradecimiento a Jorge, Helena, Alberto, Savíns y, en general, a todos aquellos que han sido partícipes de una u otra manera de la tesis. Un trocito de este éxito es sin duda vuestro.

Tampoco puedo olvidarme de mi familia: mis padres y Jose e Inma, mis hermanos. Gracias a ellos he podido llegar hasta aquí. Nunca me han negado la posibilidad de estudiar, de aprender, de trabajar y me han enseñado que la responsabilidad va por delante de todo. Cuando uno está lejos de ser brillante, como es mi caso, el trabajo y la responsabilidad son imprescindibles para conseguir lo que se propone. Así que cada día me habéis dado un empujoncito para estar aquí. Por cierto, estoy seguro de que muy pronto vamos a tener otro doctor mucho más joven y brillante en la familia: Juanjo, eres el siguiente.

Una mención especial merece Soraya, mi esposa. Ella como nadie sabe lo que me ha costado realizar este trabajo, las noches escribiendo hasta tarde y las mañanas madrugando para seguir escribiendo. Ella como nadie se ha alegrado cuando yo lo he hecho, y me ha animado cuando las cosas iban mal. Ella también ha invertido cuatro años de su vida en esta tesis, los cuatro años que no he podido dedicarle. Por eso, por estar siempre ahí y por cada día que tengo la suerte de pasar contigo: muchas gracias.

No me gustaría olvidarme de todos aquellos que, sin saberlo ni quererlo, han sido la inspiración de esta tesis. Me estoy refiriendo a todos aquellos que han tenido la enorme desgracia de tener que luchar contra una maldita enfermedad que los ha puesto en peligro. Unos han tenido la gran fortuna de poder superarla, como María (mi tía). Otros siguen luchando cada día, fuertes y hacia adelante, va por ti, Virtu. Y otros nos han dejado antes de tiempo, aunque nunca nos olvidaremos de ellos, como nuestros queridos maestros Pedro y Carmen. Todos ellos, sin darse cuenta, han sido la semilla de un trabajo que espero que pronto pueda ayudar a muchos más.

Y no, no me he olvidado, pero como lo mejor siempre se hace esperar, me gustaría dedicarle esta tesis a la persona que más orgullosa se ha sentido jamás de mí: a mi padre.

Gracias.

A Martín y Conchi, mis padres  
A Jose e Inma, mis hermanos  
A Soraya, mi mujer



## ÍNDICE GENERAL

AUTORIZACIÓN DE LOS DIRECTORES	
RESUMEN	
ABSTRACT	
AGRADECIMIENTOS	
ÍNDICE GENERAL	
SIGLAS Y ABREVIATURAS .....	17
ÍNDICE DE FIGURAS Y DE ANEXOS .....	19
GLOSARIO.....	21
<b>CAPÍTULO I: INTRODUCCIÓN.....</b>	<b>23</b>
1.1. ETAPAS EN EL PROCESO DE DISEÑO DE FÁRMACOS .....	25
1.2. TIPOS DE TÉCNICAS DE CRIBADO VIRTUAL .....	27
1.2.1. Cribado virtual basado en la estructura.....	28
1.2.2. Cribado virtual basado en ligandos.....	29
1.3. NECESIDAD DE SUPERCOMPUTACIÓN EN EL CRIBADO VIRTUAL: GRID Y CLOUD COMPUTING.....	31
1.4. LIMITACIONES Y RETOS DEL CRIBADO VIRTUAL .....	32
<b>CAPÍTULO II: OBJETIVOS .....</b>	<b>35</b>
<b>CAPÍTULO III: ARTÍCULOS QUE COMPONEN LA TESIS DOCTORAL ..</b>	<b>39</b>
3.1. FUNDAMENTACIÓN TEÓRICA DE LAS PUBLICACIONES .....	41
3.2. ADVANCES IN DISTRIBUTED COMPUTING WITH MODERN DRUG DISCOVERY .....	43
3.3. A REVIEW OF LIGAND-BASED VIRTUAL SCREENING WEB TOOLS AND SCREENING ALGORITHMS IN LARGE MOLECULAR DATABASES IN THE AGE OF BIG DATA .....	58
3.4. BRUSELAS: HPC GENERIC AND CUSTOMIZABLE SOFTWARE ARCHITECTURE FOR 3D LIGAND-BASED VIRTUAL SCREENING OF LARGE MOLECULAR DATABASES.....	77
<b>CAPÍTULO IV: CONCLUSIONES .....</b>	<b>91</b>
4.1. CONCLUSIONES.....	93
4.2. FUTURAS LÍNEAS DE INVESTIGACIÓN .....	95

<b>CAPÍTULO V: REFERENCIAS BIBLIOGRÁFICAS .....</b>	<b>97</b>
<b>CAPÍTULO VI: ANEXOS .....</b>	<b>107</b>
6.1. CALIDAD DE LAS PUBLICACIONES.....	109
6.2. OTRAS PUBLICACIONES .....	115
6.3. PROYECTOS DE INVESTIGACIÓN PARTICIPADOS.....	118



## SIGLAS Y ABREVIATURAS

<b>ADME-T</b>	Absortion, Distribution, Metabolism, Excretion and Toxicity
<b>BRUSELAS</b>	Balanced, Rapid and Unrestricted Server for Extensive Ligand-Aimed Screening
<b>CADD</b>	Computer-Aided Drug Design
<b>GPU</b>	Graphics Processing Unit
<b>HPC</b>	High-Performance Computing
<b>LBVS</b>	Ligand-Based Virtual Screening
<b>QSAR</b>	Quantitative Structure-Activity Relationship
<b>RMSD</b>	Root-Mean-Squeared Deviation
<b>SBVS</b>	Structure-Based Virtual Screening
<b>VS</b>	Virtual Screening



## ÍNDICE DE FIGURAS, Y DE ANEXOS

### ÍNDICE DE DE FIGURAS

Figura 1.1: Proceso tradicional de descubrimiento de fármacos .....	26
Figura 1.2: Clasificación de las técnicas de cribado virtual .....	28
Figura 1.3: Técnicas de cribado virtual en función de su precisión, velocidad y capacidad de cómputo requerida .....	30
Figura 1.4: Tipos más comunes de computación distribuida .....	31
Figura 6.1: Datos identificativos de Expert Opinion on Drug Discovery .....	109
Figura 6.2: Evolución del factor de impacto y del percentil en su categoría de Expert Opinion on Drug Discovery .....	110
Figura 6.3: Indicadores clave de Expert Opinion on Drug Discovery .....	110
Figura 6.4: Datos identificativos de Future Medicinal Chemistry .....	111
Figura 6.5: Evolución del factor de impacto y del percentil en su categoría de Future Medicinal Chemistry .....	111
Figura 6.6: Indicadores clave de Future Medicinal Chemistry .....	112
Figura 6.7: Datos identificativos de Journal of Chemical Information and Modeling .....	113
Figura 6.8: Evolución del factor de impacto y del percentil en su categoría de Journal of Chemical Information and Modeling .....	113
Figura 6.9: Indicadores clave de Journal of Chemical Information and Modeling .....	114
Figura 6.10: Portada de Journal of Chemical Information and Modeling .....	117

**ÍNDICE DE ANEXOS**

Anexo 1: Calidad de las publicaciones.....	109
Anexo 2: Otras publicaciones .....	115
Anexo 3: Proyectos de investigación participados .....	118

## GLOSARIO

**Compuesto candidato (*lead*):** es una molécula que ha mostrado propiedades que la hacen ser candidata a interactuar con un cierto receptor. Cuando el *lead* procede de un proceso de cribado virtual, debe haber sido sintetizado y optimizado para mejorar su actividad biológica.

**Conformación:** es cada una de las disposiciones espaciales que puede adoptar una molécula como consecuencia de la libertad de rotación en torno a enlaces sigma en ciertas partes de su estructura.

**Descriptor:** es una propiedad molecular (peso, número de átomos...). Su valor es habitualmente calculado mediante software (p.ej. DRAGON, CDK, RDKit).

**Ligando:** es una pequeña molécula que se une a una macromolécula (p.ej. proteína, ADN) ya sea mediante interacciones no covalentes (p.ej. enlaces por puentes de hidrógeno) o covalentes. La unión se produce en una parte específica de la macromolécula, ya sea en el sitio activo de una enzima, o en un sitio alostérico cualquiera.

**Pose:** es cada una de las posibles posiciones de una conformación en el espacio tridimensional mediante rotaciones y traslaciones.

**Receptor:** en una macromolécula (normalmente una proteína) con la que interactúa un ligando. Al recibir la señal del ligando, el receptor sufre un cambio ya sea en su forma o en su actividad.

**Sitio de unión:** es la región del receptor a la que un ligando puede unirse ya sea mediante interacciones covalentes, y por tanto irreversibles, o no covalentes.



# **I - INTRODUCCIÓN**





## I - INTRODUCCIÓN

En la actualidad, millones de personas en todo el mundo sufren enfermedades o trastornos para los que aún no existe un tratamiento efectivo [1-3]. Pese a los avances logrados en las últimas décadas, todavía no se dispone de fármacos ni tratamientos eficaces contra algunas de las dolencias más frecuentes de nuestros días, como son el Alzheimer, el Parkinson o el cáncer [4-6]. Con el ánimo de avanzar más rápida y eficazmente en la búsqueda de fármacos efectivos contra estos problemas de salud, ciertos campos como la biología, la química, las matemáticas, la medicina y la informática se han unido en una nueva rama interdisciplinaria llamada bioinformática. El objetivo de la bioinformática es aprovechar las técnicas de computación de alto rendimiento (*high performance computing*, HPC) para resolver cálculos complejos sobre grandes volúmenes de datos en un corto espacio de tiempo, por ejemplo, el cribado de grandes bases de datos moleculares. Esta nueva ciencia ya ha dado frutos importantes, como el descubrimiento de nuevos fármacos y el desarrollo de terapias celulares [7-9]. Precisamente el diseño de fármacos asistido por computador (*computer-aided drug design*, CADD) es uno de los campos más frecuentes en los que se aplica la bioinformática, ya que, en la era del big data en la que vivimos, hay una abrumadora cantidad de compuestos químicos a analizar y el uso de HPC ayuda a reducir drásticamente el tiempo necesario para ello [10].

### 1.1 ETAPAS EN EL PROCESO DE DISEÑO DE FÁRMACOS

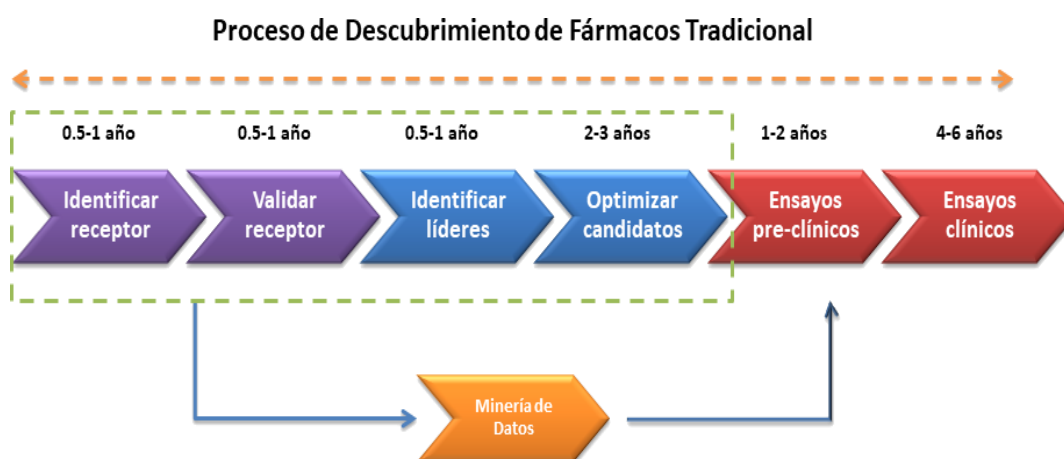
La duración del proceso de desarrollo de un nuevo fármaco se estima en más de 10 años y un coste económico que, según algunos estudios, supera los 2 billones de dólares [11, 12].

El proceso tradicional, mostrado en la Figura 1, comienza con la identificación y validación de la macromolécula receptora (habitualmente una proteína) que interactúa con otra macromolécula o con una pequeña molécula, llamada ligando, provocando una reacción indeseada que ocasiona la enfermedad en cuestión. En muchas ocasiones no es posible determinar cuál es el receptor que

interviene en la anomalía, y se debe comenzar a investigar a partir del ligando con el que reacciona o de un conjunto de ligandos de los que se posee cierta información sobre su actividad biológica. Conocer el receptor de una reacción biológica, como se explica más adelante, permite aplicar técnicas más avanzadas a la hora de encontrar fármacos potenciales. Sin embargo, no siempre se dispone de esta información, en cuyo caso se deben emplear métodos alternativos.

A continuación se busca una colección de compuestos líderes, es decir, ligandos cuyas propiedades muestran que son candidatos a interactuar con el receptor. Una vez identificados los *leads* con los que se va a trabajar, estos deben ser cuidadosamente validados y optimizados con el fin de predecir con la mayor certeza posible si podrán desempeñar la función deseada. Por ejemplo, el estudio de las propiedades ADME-T (*Absorption, Distribution, Metabolism, Excretion and Toxicity*) puede determinar si las condiciones de absorción o toxicidad del ligando son las adecuadas para su funcionamiento como fármaco.

Finalmente, las etapas más costosas, en términos del tiempo empleado, son los ensayos pre-clínicos y clínicos. Estas fases pueden prologarse hasta 8 años, y son especialmente delicadas puesto que son las que permiten determinar con exactitud si los potenciales fármacos son realmente efectivos y cuáles son sus efectos secundarios.



**Figura 1.1.** Proceso tradicional de descubrimiento de fármacos.

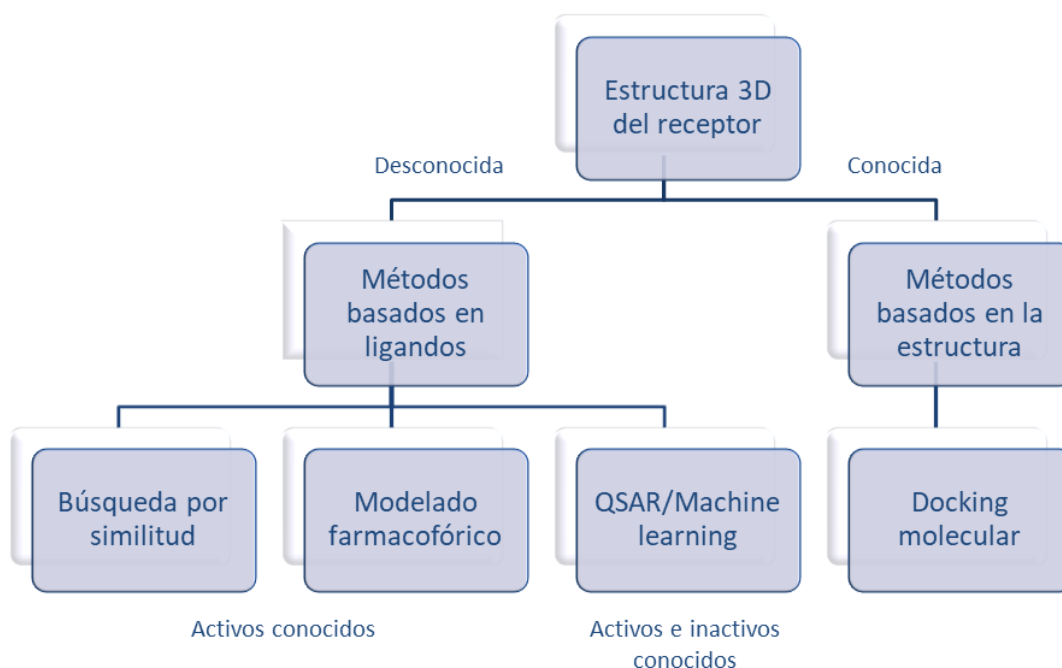
En las etapas iniciales nos encontramos con el problema añadido del enorme tamaño del espacio químico disponible debido, en parte, a la aparición de

bases de datos químicas de libre acceso que almacenan desde miles hasta millones de compuestos (p.ej. PubChem [13], GDB-17 [14]). En un estudio reciente, Hoffman et al. [15] estiman el número de compuestos disponibles, calculado a partir de datos obtenidos de la industria farmacéutica, en  $10^{20}$ . Sin embargo, Polischuk et al. [16] van más allá y hablan de  $10^{33}$  compuestos disponibles a partir de la base de datos GDB-17. Estas cifras, que solamente consideran fármacos potenciales, hacen totalmente inviable el método de ensayo y error para determinar qué compuestos son reactivos con el receptor deseado.

Con objeto de reducir este espacio químico a uno mucho más manejable, es habitual el uso de técnicas computacionales en las primeras fases del proceso de desarrollo de fármacos, especialmente en la etapa de búsqueda de líderes. Esta reducción del espacio químico mediante métodos *in silico* se conoce como cribado virtual o *virtual screening* (VS).

## 1.2 TIPOS DE TÉCNICAS DE CRIBADO VIRTUAL

El VS es una técnica computacional que permite buscar en grandes librerías de moléculas, aquellas estructuras que tienen más probabilidades de enlazarse a un receptor [17]. Existen diversos tipos de VS y la aplicación de uno u otro dependerá de la información disponible en cada caso. Tradicionalmente, los métodos de VS se clasifican en dos grupos principales: basados en la estructura (*structure-based virtual screening*, SBVS) y basados en ligandos (*ligand-based virtual screening*, LBVS). La distinción entre unos y otros estriba en la información disponible acerca de la estructura tridimensional del receptor (Figura 1.2).



**Figura 1.2.** Clasificación de las técnicas de cribado virtual.

### 1.2.1 Cribado virtual basado en la estructura

Cuando se conoce cuál es la estructura tridimensional del receptor, obtenida por métodos como la cristalografía, los rayos X o la resonancia magnética nuclear, se suelen aplicar en primer lugar las técnicas SBVS. Entre ellas, probablemente la más típica sea el acoplamiento molecular o *molecular docking* (o simplemente *docking*). El acoplamiento molecular permite encontrar el modo de unión más favorable de un ligando con un receptor para formar un complejo estable. Esta técnica es muy utilizada por su habilidad para predecir, con una alta precisión, la pose más adecuada del ligando en un sitio de unión determinado del receptor. Para ello, se evalúa la energía requerida por cada pose para interactuar con el receptor en un sitio de unión dado, y se repite el proceso hasta alcanzar la pose de menor energía entre todas las calculadas sobre la superficie del receptor.

Aunque el acoplamiento molecular es uno de los tipos de métodos SBVS más frecuentemente utilizados, no es el único. Dentro de este grupo de técnicas,

también se encuentran las simulaciones de dinámica molecular (*molecular dynamics*, MD), en las que se realiza una simulación del movimiento de los átomos y las moléculas durante un período de tiempo. De este modo se consigue visualizar la posición y velocidad de cada átomo para cada instante. Las técnicas SBVS también pueden refinarse incluyendo la descripción más precisa que los cálculos mecánico-cuánticos (*quantum mechanical calculations*, QM) hacen de las interacciones a nivel molecular, incluyendo la posible formación de enlaces químicos entre receptor y ligando. Sin embargo, los cálculos mecánico-cuánticos tienen un coste computacional mucho más elevado que los métodos anteriormente comentados.

En general, las técnicas SBVS son las más utilizadas cuando se dispone de la estructura 3D del receptor [18, 19]. Desafortunadamente, existen dos problemas importantes a la hora de aplicarlas: i) no siempre se conoce la estructura 3D del receptor; y ii) los complejos cálculos que realizan (p.ej. Funciones consenso) las hacen ser computacionalmente muy costosas.

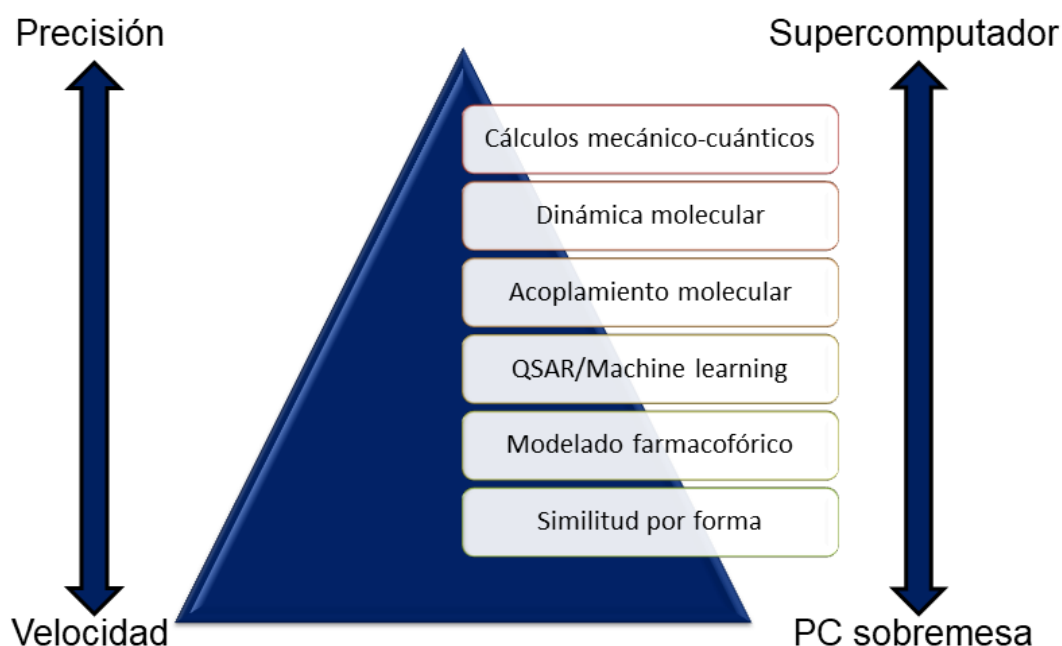
### 1.2.2 Cribado virtual basado en ligandos

En el caso de no poder aplicar métodos SBVS o que su cálculo sea demasiado costoso, los métodos basados en ligandos son la alternativa preferida. Este grupo de técnicas son, de manera general, una alternativa eficiente a las SBVS porque, aunque sean algo menos precisas en ciertos casos, su coste computacional es mucho menor [20]. Debido a esta eficiencia para obtener los resultados, a menudo se emplean como un paso previo al SBVS para obtener un reducido conjunto de compuestos candidatos que, posteriormente, pueda ser procesado por métodos más costosos y precisos (Figura 1.3).

El grupo de LBVS incluye tres técnicas fundamentalmente: la búsqueda por similitud, el modelado farmacofórico y las relaciones cuantitativas estructura-actividad (*quantitative structure-activity relationship*, QSAR). Todas ellas parten del supuesto de que los compuestos estructuralmente parecidos muestran una actividad biológica similar [21].

Cuando el conjunto de ligandos de partida contiene únicamente información acerca de moléculas inactivas, entonces es común optar por las dos primeras. La búsqueda por similitud compara directamente la estructura

molecular de los ligandos. Al ser una técnica relativamente sencilla con un coste computacional bajo, es muy atractiva para cribar grandes bases de datos [22]. El modelado farmacofórico pretende detectar, de forma sistemática, una colección de características que sean las mínimas necesarias para garantizar una interacción óptima con el receptor. Sólo aquellos ligandos cuyas características se adapten mejor al modelo serán considerados como resultados aceptables.



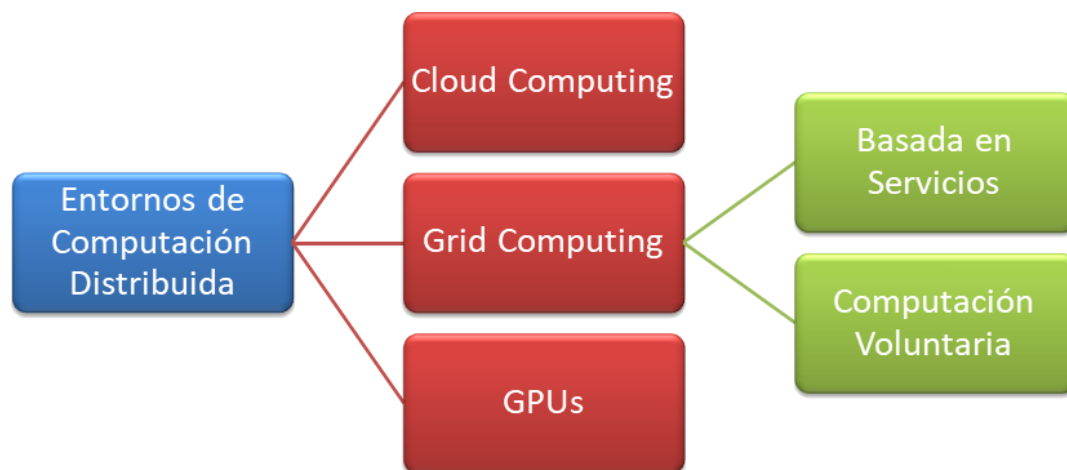
**Figura 1.3.** Técnicas de cribado virtual en función de su precisión, velocidad y capacidad de cómputo requerida.

Si, por el contrario, se dispone de un conjunto de ligandos activos e inactivos con el receptor, entonces se pueden aplicar modelos matemáticos para establecer una relación entre la estructura molecular y la actividad biológica asociada. Los modelos matemáticos utilizados pueden ser tanto lineales como no lineales. Los primeros incluyen las técnicas QSAR, que aplican modelos matemáticos y estadísticos, como la regresión lineal, para predecir la actividad de la molécula. Por su parte, los modelos no lineales engloban las técnicas de aprendizaje automático o *machine learning*, en las que el sistema realiza sus

predicciones en base a lo aprendido a través de un conjunto de ejemplos durante una fase previa de entrenamiento [23].

### 1.3 NECESIDAD DE SUPERCOMPUTACIÓN EN EL CRIBADO VIRTUAL: GRID Y CLOUD COMPUTING

La gran cantidad de datos disponibles y la complejidad de los cálculos hacen que el VS sea un proceso computacionalmente muy costoso [24]. La potencia computacional y la memoria disponibles en una estación de trabajo convencional no son suficientes para ejecutar millones de complejas operaciones en un tiempo admisible, por lo que es necesario poner al servicio de VS, y de CADD en general, técnicas avanzadas de supercomputación. Los enfoques *grid* y *cloud computing* ya se han aplicado a VS, consiguiendo reducir el tiempo de cálculo de manera drástica [25], y se están convirtiendo en una vía eficaz para mejorar el rendimiento de este tipo de procesos. La Figura 1.4 muestra la clasificación típica de los entornos de computación distribuida.



**Figura 1.4.** Tipos más comunes de computación distribuida.

El paradigma *grid computing* consiste en distribuir las tareas entre distintos nodos geográficamente dispersos que ejecutan procesos independientes. En las últimas décadas ha sido muy utilizado y se ha aplicado con éxito en el uso del

acoplamiento molecular y las dinámicas moleculares para cribar librerías de hasta 600.000 ligandos [26, 27]. Este tipo de técnicas SBVS suelen trabajar con pequeñas librerías puesto que requieren mucha capacidad de cálculo, sin embargo, el uso de *grid computing* ha permitido utilizarlas con librerías más grandes. Pese a esto, la complejidad creciente de los cálculos en VS, la necesidad de mantener una configuración compatible en todos los nodos y la dificultad para compartir información entre ellos, han llevado al uso progresivo de *cloud computing* [28].

La computación en la nube o *cloud computing*, por el contrario, permite ejecutar tareas en entornos de computación virtualizados que pueden ser fácilmente escalados bajo demanda y, con la ventaja adicional, de consumir menos energía [29]. Bajo este paradigma, Capuccini et al. [30] consiguieron acelerar un 87% la duración de una tarea de *docking*, lo que anima a ser optimistas de cara a un futuro cercano.

Otros enfoques como la computación voluntaria (*volunteer computing*) y las GPUs (*graphics processing units*) están creciendo en importancia en el campo del VS en los últimos años. La computación voluntaria es un tipo de *grid computing* donde cada nodo ejecuta sus tareas en segundo plano de manera colaborativa pudiendo conseguir una potencia de cálculo y de almacenamiento comparable a la de algunos clusters [31]. Este modelo ya ha sido aplicado con éxito en CADD [32] y está emergiendo como una buena opción para resolver determinados problemas bioinformáticos, abriendo así una nueva puerta a futuros estudios [33]. Por su parte, la tecnología GPU es un tipo de computación paralela que ha evolucionado desde su uso en videojuegos hasta emplearse para acelerar aplicaciones no gráficas. En este sentido, las GPUs han sido capaces de reducir drásticamente el tiempo de cálculo en problemas de cribado virtual con redes neuronales, cálculos mecánico-cuánticos y búsquedas por similitud [34-36], y se espera que el número de aplicaciones basadas en esta tecnología continúe creciendo.

#### 1.4 LIMITACIONES Y RETOS DEL CRIBADO VIRTUAL

Gracias al uso de la supercomputación, el VS ha conseguido alcanzar una buena relación entre la calidad de los resultados y el tiempo invertido en



obtenerlos, lo que lo hace una alternativa muy atractiva para el descubrimiento de fármacos. Sin embargo, pese a la ayuda de la supercomputación, aún quedan algunas limitaciones y retos por superar.

Como ya se ha mencionado, el número de compuestos crece a una velocidad vertiginosa [37], lo que invita al uso de técnicas LBVS, como paso previo a otros métodos más costosos, debido a su rapidez y eficiencia. El primer paso será, por tanto, la elección de un método LBVS y un software que lo implemente. Se pueden encontrar numerosos paquetes software que implementan estos métodos, ya sean comerciales, de licencia académica o de libre acceso, como por ejemplo USR [38], USRCAT [39] y ElectroShape [40]. Una primera limitación a la hora de utilizarlos es que son necesarios ciertos conocimientos para su instalación y uso, que no todos los usuarios poseen.

La alternativa para evitar la instalación de este tipo de software es su utilización desde una máquina remota. Esto ha dado lugar a la aparición de servidores web que prestan servicios de VS, como SwissSimilarity [41], ChemMapper [42] y ZINCPharmer [43]. Estos servidores proporcionan una interfaz más sencilla, que oculta la compleja configuración del software subyacente. Habitualmente los usuarios pueden elegir unos pocos parámetros para ejecutar sus tareas y el servidor se encarga del resto. Entre esos parámetros se encuentran las librerías a cribar y el algoritmo de cribado a utilizar. En cuanto a las librerías disponibles para VS, la tendencia es procesar librerías pequeñas y muy focalizadas en un problema concreto. El número de librerías disponibles en los servidores actuales suele ser limitado y fijo para evitar que las tareas de cribado tarden mucho tiempo e impidan el acceso a los recursos por parte de otros usuarios. Sin embargo, este planteamiento limita mucho a los usuarios cuando lo que se desea es trabajar con una librería preparada *offline* o el cribado de un conjunto de compuestos muy concreto.

El otro parámetro relevante de todo servidor web de VS es el algoritmo que emplea para evaluar la similitud molecular. Tradicionalmente, los servidores de VS simplemente servían de interfaz gráfica de los algoritmos implementados, de manera que los usuarios tenían acceso al algoritmo sin necesidad de instalarlo. Poco a poco, la tendencia fue variando, y muchos de los servidores actuales permiten elegir uno de entre un conjunto de algoritmos. Por desgracia, cada

algoritmo evalúa la similitud molecular de una manera distinta, lo que hace que, para el mismo par de moléculas, dos algoritmos puedan medir grados de similitud diferentes. Estas diferencias se suelen traducir en un valor numérico variante, como el coeficiente de Tanimoto o el RMSD, que hace que los resultados sean inconsistentes y difíciles de interpretar. Además, no todos los algoritmos muestran las mismas prestaciones en todos los casos, pudiendo llevar a resultados sesgados en función del algoritmo utilizado en cada tarea. Una manera de solventar esta dependencia de un único algoritmo, sería la combinación de varios de ellos en un mismo proceso de VS. En consecuencia, sería necesario aplicar una función de consenso para combinar sus puntuaciones y corregir las desviaciones que puedan tener los algoritmos individuales.

Por lo tanto, considerando todas las limitaciones expuestas, parece necesario el desarrollo de un nuevo servidor de LBVS que sea capaz de crear librerías de compuestos de manera dinámica y combinar varios algoritmos de similitud en una misma tarea para evitar sesgos. La aplicación del consenso de las puntuaciones aumentará la precisión de las predicciones con respecto a los algoritmos individuales. Es evidente, también, la necesidad de apoyarse en alguna plataforma HPC para que los resultados puedan estar disponibles en un tiempo admisible. Por ejemplo, utilizando un clúster de computación paralela se podría reducir de manera drástica el tiempo de cálculo necesario para procesar las librerías. Y, al tratarse de un servidor web, será accesible desde cualquier dispositivo y evitará a los usuarios la necesidad de instalar y configurar manualmente todo el software utilizado.

## **II - OBJETIVOS**



## II - OBJETIVOS

Al comienzo de esta tesis se plantearon una serie de objetivos principales a cumplir que se describen a continuación. A medida que la investigación fue avanzando, algunos de los objetivos se fueron refinando para cubrir más casos de uso.

**Objetivo 1.** Hacer una revisión profunda de los servidores de LBVS existentes en la actualidad. Se deben identificar los parámetros de entrada que requieren, los algoritmos de similitud que utilizan, las bases de datos que criban y el tiempo estimado de cómputo que consumen. La revisión se completará con un estudio de las técnicas HPC que se emplean frecuentemente en CADD, detallando sus ventajas e inconvenientes.

**Objetivo 2.** El desarrollo de una herramienta web que facilite la ejecución de tareas de LBVS a los usuarios a través de una interfaz web amigable. La herramienta debe ser de aplicabilidad general, es decir, que pueda ser usada para buscar compuestos que se puedan aplicar en cualquier contexto, no únicamente el farmacológico. Además, se desea importar un gran número de compuestos de bases de datos de libre acceso, y generar distintas conformaciones de cada uno para crear una base de datos propia del servidor. Los compuestos serán importados desde DrugBank [44], ChEMBL [45], KEGG [46] y DIA-DB [47]. Las dos primeras son bases de datos de libre acceso de las que se puede extraer abundante información sobre la actividad biológica de los compuestos. Por su parte, de KEGG se importarán los subconjuntos Drugs y Compounds, que permitirán completar la colección de candidatos a fármacos y añadirán nuevos compuestos para otro tipo de contextos. DIA-DB es una pequeña base de datos de antidiabéticos creada *in house*. Los usuarios también podrán someter sus propias bases de datos a cribado. Para que el rendimiento del servidor no se vea afectado por la cantidad de moléculas a cribar, se apoyará en una plataforma HPC donde ejecutará los cálculos en paralelo. En concreto, se utilizará un clúster de

supercomputación ubicado en el Laboratorio Nacional para Computación de Alto Rendimiento de Chile (*National Laboratory for High Performance Computing, NLHPC*), el cual permite utilizar hasta 15 nodos de computación a la vez.

**Objetivo 3.** Se debe generar un valor científico añadido, implementando funcionalidades no existentes hasta la fecha en otras herramientas similares. Las nuevas características deben incluir, al menos:

- El uso de varios algoritmos de similitud en una misma tarea y la aplicación de una función de consenso para combinar sus puntuaciones.
- Crear librerías de compuestos a partir de un conjunto de descriptores seleccionados por el usuario. Todos los descriptores serán calculados con el software DRAGON [48] para estandarizar los valores.
- Implementar búsqueda por similitud y cribado farmacofórico. Los algoritmos incluidos serán LiSiCA [49], Screen3D [50], WEGA [51] y OptiPharm [52] para la similaridad, y SHAFTS [53] y Pharmer [54] para el cribado farmacofórico.
- Mostrar gráficamente cómo se alinean la molécula de referencia y cada uno de los ligandos con cada algoritmo elegido. Los alineamientos también deben poder ser procesados *offline*, por lo que todos los ligandos alineados y las poses de la molécula de entrada deben poder descargarse para ser manipulados con PyMOL. También sería muy interesante descargar las puntuaciones calculadas en un fichero Excel.
- Importar información textual de los compuestos importados. Si los usuarios lo desean, el cribado considerará sólo los compuestos relacionados con ciertos términos clave, por ejemplo, el nombre de una enfermedad o de un receptor.

**Objetivo 4.** Una vez desarrolladas las nuevas funcionalidades, habrá que aplicarlas a casos de estudio prácticos. Por ejemplo, la validación de la herramienta comparando sus predicciones con casos reales o proponer, de manera teórica, un conjunto de fármacos potenciales para tratar una determinada enfermedad.

**III – ARTÍCULOS QUE  
COMPONEN LA TESIS  
DOCTORAL**





### III – ARTÍCULOS QUE COMPONEN LA TESIS DOCTORAL

#### 3.1 FUNDAMENTACIÓN TEÓRICA DE LAS PUBLICACIONES

El núcleo de esta tesis doctoral se desglosa a lo largo de los 3 artículos que componen el compendio, los cuales han sido publicados en revistas indexadas de alto impacto en el orden en que son presentados en este capítulo. La tesis ha de ser abordada tanto desde el punto de vista computacional como del bioquímico. Con esta idea en mente, la división de los artículos atiende a la necesidad de aplicar técnicas de supercomputación en VS, analizar las herramientas web de LBVS existentes y desarrollar de una nueva herramienta que cubra los aspectos no explorados hasta el momento. De este modo se cubren ambos puntos de vista, los cuales se van entrelazando en cada publicación.

Como se ha explicado en la introducción, la necesidad de trabajar con una extraordinariamente alta cantidad de datos hace que la supercomputación sea imprescindible en el desarrollo de fármacos y, más concretamente, en el cribado virtual. Por este motivo, el primer artículo describe en detalle las técnicas HPC más frecuentes para acelerar el rendimiento de aplicaciones bioinformáticas. Se hace una profunda exploración de los paradigmas *grid* y *cloud computing*, y del uso de GPUs y de la computación distribuida en la bioinformática actual. El artículo detalla las principales similitudes y diferencias entre los distintos paradigmas, así como sus ventajas y limitaciones. Además, presenta diversos ejemplos de herramientas de VS que se apoyan en plataformas HPC para mejorar sus prestaciones.

Puesto que los métodos de LBVS son una alternativa eficiente para cribar grandes bases de datos químicas, el segundo artículo se centra en los servidores web que implementan este tipo de procesos. Una exhaustiva revisión de estos servidores a lo largo de su historia muestra cuántos y qué algoritmos soporta cada servidor, qué bases de datos son capaces de cribar, las funcionalidades más destacadas y el rendimiento de cada uno en casos similares. El estudio no se limita a las herramientas web disponibles, sino que también analiza los

algoritmos de similitud molecular subyacentes y otros aspectos que los pueden condicionar, como la representación molecular empleada, los descriptores moleculares asociados a cada tipo de representación y los filtros moleculares más típicos. Como conclusión de esta revisión se identifican los puntos fuertes y las debilidades de los servidores actuales.

Una vez identificadas las fortalezas y debilidades de las herramientas actuales, se presenta la arquitectura BRUSELAS, cuyo objetivo es proporcionar, de manera gratuita, una solución web que aporte un valor añadido a los servidores estudiados. El aporte novedoso de BRUSELAS lo componen, principalmente, la combinación de algoritmos de similitud mediante funciones de consenso y el cribado de grandes librerías mediante el uso de descriptores moleculares. Adicionalmente, también permite focalizar las librerías generadas por medio de palabras clave y limitar los resultados aplicando el filtro de Lipinski. Cabe recordar que la regla del 5 de Lipinski es un filtro molecular frecuentemente utilizado, que utiliza las propiedades farmacocinéticas de un compuesto candidato para decidir si es lo bastante tóxico como para impedir su uso como fármaco [55].

Para alcanzar un rendimiento competitivo con el resto de servidores, BRUSELAS hace uso de un clúster de supercomputación que le permite obtener los resultados en tiempo similar o menor al de sus competidores. Este último artículo se completa con una comparativa del rendimiento y la fiabilidad de las predicciones realizadas por BRUSELAS frente a un conjunto representativo de los servidores previamente introducidos. Además, se utiliza la nueva arquitectura para proponer compuestos candidatos a fármacos en algunos ámbitos concretos, como son la búsqueda de anticoagulantes sanguíneos o las terapias contra el cáncer.

## 3.2 ADVANCES IN DISTRIBUTED COMPUTING WITH MODERN DRUG DISCOVERY

<b>Título</b>	Advances in Distributed Computing With Modern Drug Discovery
<b>Autores</b>	Banegas-Luna AJ, Imbernón B, Llanes Castro A, Pérez-Garrido A, Cerón-Carrasco JP, Gesing S, Merelli I, D'Agostino D, Pérez-Sánchez H.
<b>Revista</b>	Expert Opinion on Drug Discovery
<b>Año</b>	2018
<b>Volumen</b>	14 (1)
<b>Páginas</b>	9-22
<b>Estado</b>	Publicado
<b>DOI</b>	<a href="https://doi.org/10.1080/17460441.2019.1552936">https://doi.org/10.1080/17460441.2019.1552936</a>
<b>IF(2017)</b>	4.692
<b>Categoría</b>	Pharmacology & Pharmacy, 23/261, Q1(D1)

**Contribución del Doctorando**

El doctorando Antonio Jesús Banegas Luna declara ser el autor principal y contribuyente principal del artículo *Advances in Distributed Computing with Modern Drug Discovery*.

## REVIEW



## Advances in distributed computing with modern drug discovery

Antonio Jesús Banegas-Luna<sup>a</sup>, Baldomero Imbernón<sup>a</sup>, Antonio Llanes Castro <sup>a</sup>, Alfonso Pérez-Garrido<sup>a</sup>, José Pedro Cerón-Carrasco<sup>a</sup>, Sandra Gesing<sup>b</sup>, Ivan Merelli<sup>c</sup>, Daniele D'Agostino<sup>d</sup> and Horacio Pérez-Sánchez<sup>a</sup>

<sup>a</sup>Bioinformatics and High Performance Computing Research Group (BIO-HPC), Universidad Católica de Murcia (UCAM), Murcia, Spain; <sup>b</sup>Center for Research Computing, University of Notre Dame, Notre Dame, IN, USA; <sup>c</sup>Institute for Biomedical Technologies, National Research Council of Italy, Segrate (Milan), Italy; <sup>d</sup>Institute for Applied Mathematics and Information Technologies “E. Magenes”, National Research Council of Italy, Genoa, Italy

### ABSTRACT

**Introduction:** Computational chemistry dramatically accelerates the drug discovery process and high-performance computing (HPC) can be used to speed up the most expensive calculations. Supporting a local HPC infrastructure is both costly and time-consuming, and, therefore, many research groups are moving from in-house solutions to remote-distributed computing platforms.

**Areas covered:** The authors focus on the use of distributed technologies, solutions, and infrastructures to gain access to HPC capabilities, software tools, and datasets to run the complex simulations required in computational drug discovery (CDD).

**Expert opinion:** The use of computational tools can decrease the time to market of new drugs. HPC has a crucial role in handling the complex algorithms and large volumes of data required to achieve specificity and avoid undesirable side-effects. Distributed computing environments have clear advantages over in-house solutions in terms of cost and sustainability. The use of infrastructures relying on virtualization reduces set-up costs. Distributed computing resources can be difficult to access, although web-based solutions are becoming increasingly available. There is a trade-off between cost-effectiveness and accessibility in using on-demand computing resources rather than free/academic resources. Graphics processing unit computing, with its outstanding parallel computing power, is becoming increasingly important.

### ARTICLE HISTORY

Received 28 February 2018  
Accepted 23 November 2018

### KEYWORDS

Cloud computing;  
computational chemistry;  
distributed computing; drug  
discovery; grid computing;  
high-performance  
computing; virtual screening

## 1. Introduction

It is now widely accepted that the discovery of new drugs can be aided by the use of computational drug discovery (CDD) techniques and many approved drugs have reached and passed clinical trials with their help. New terms have been added to the vocabulary of researchers, including computer-aided drug design (CADD) [1], computer-aided molecular design (CAMD) [2] and computer-aided molecular modeling (CAMM).

Most drug discovery studies focus on enhancing specific parts of the processes involved in the development of new drugs, which can be divided into three phases: (1) a discovery phase, in which millions of candidate compounds are screened; (2) a selection phase, in which the candidate drugs undergo preclinical research; and (3) an assessment phase, in which the drug is developed and extensive clinical trials conducted. *In silico* solutions are usually carried out in the discovery phase, whereas screening was previously carried out in a laboratory over several years with significant economic costs.

Many computational techniques are now available to study the molecular interactions relevant to drug discovery, such as virtual screening (VS) [3], which is used to simulate a large number of interactions between proteins (also known as receptors and/or enzymes) and small molecule drug candidates (ligands). Docking software is usually tested on protein families where there are the

most crystal structures and, therefore, it is common practice to test many proteins in parallel [4–6]. Side-effects caused by off-target bindings should be avoided and, therefore, the most promising compounds are usually tested against many other proteins. The docking conformations that describe the interactions between each compound and the corresponding target are optimized through molecular dynamics (MD) simulations to relax the system and improve the accuracy with which the binding energy is calculated. MD is a physics-based simulation method in which Newton's equations of motion are solved for each atom of the system considering all the forces involved in their interactions [7]. Depending on the number of atoms involved, it can be computationally demanding. Other cheaper techniques, such as chemical similarity and calculating the proximity matrix, are also used in the drug discovery process [8–10].

CDD methods can be used in either a predictive or a prospective way. In predictive CDD, calculations are carried out to process a database of compounds and to anticipate which compounds are most likely to be of interest before their characterization in the laboratory. In prospective CDD, the experimental results obtained after screening compounds in the laboratory are analyzed by CDD methods to try to understand why the compounds were selected. In both these approaches, the complexity of the computations depends on the size of the database and/or the accuracy of the methods used. The use of high-performance

**Article highlights**

- VS is a computational chemistry approach that is currently used to reduce the number of wet laboratory experiments required in drug discovery campaigns. However, it may require a considerable amount of computing power, the availability of which is an issue for many research groups and small- or medium-sized companies.
- Distributed computing is a cost-effective solution in drug discovery research and can be implemented in a number of ways. Grid computing is the most efficient way of distributing the demand involved in calculations across a network of available computing units. Cloud computing is deployed through virtual hardware resources and is a more flexible approach than grid computing because it can be configured and scaled depending on the complexity of the task.
- Within distributed computing platforms, graphics processing unit computing increasingly has a key role as a result of the parallel nature of the hardware, which increases the throughput in most scientific computing tasks. This can be useful for the computationally demanding tasks involved in drug discovery. However, not all computational chemistry codes are amenable to efficient parallelization.
- Cloud applications are now available to support the deployment of simulations to non-expert users at an early stage. Web servers are available for VS calculations, which implement many different CDD techniques, although most of these are for ligand-based methods because these are computationally cheaper than structure-based methods.
- Even when using distributed computing approaches, the size of the chemical space is still too large (millions of compounds) to be covered by one single VS technique. Therefore a hierarchical VS approach is often adopted: inexpensive methods (e.g. similarity searching or pharmacophore modeling) are applied first to create small, focused libraries, followed by more computationally expensive methods (e.g. molecular docking and MD) to achieve the best results.
- Structure-based VS approaches, such as molecular docking or MD, generally require more computational resources than ligand-based methods. However, ligand-based techniques are less accurate than structure-based methods and, therefore, a trade-off is required when performing computations in distributed computing platforms to obtain the best results according to the resources available.

This box summarizes key points contained in the article.

computing (HPC) techniques is now mandatory in the development of efficient and scalable tools for CDD. Many different HPC approaches can be used, including graphics processing units (GPUs), in-house clusters of computers, remote supercomputers and distributed computing infrastructures.

The creation and maintenance of a local computing facility capable of managing the huge amount of data obtained from state-of-the-art acquisition instruments and simulation tools in drug discovery projects is too expensive for many small- or medium-sized biotechnology laboratories [11] and the use of remote computational services is a cost-effective solution. Remote computing infrastructures provide researchers with the ability to adjust the computational resources according to their actual requirements, whereas doubling the size of a traditional cluster is expensive and can result in many idle resources during off-peak periods. In a distributed environment, requests to double the amount of access to resources simply involve paying for twice the capacity. Usability is the most crucial aspect in this scenario because end-users require solutions that simplify their activities rather than making them more complex.

We review here the current trends and advances in the application of distributed computing infrastructures to drug discovery. Section 2 reviews the most commonly used tools in

ligand- or structure-based VS. The tools discussed here provide a diversity of screening services, such as quantitative structure–activity relationship (QSAR) modeling, docking and molecular docking, and MD. Section 3 introduces the currently available distributed computing environments, including advances in the use of GPU-based HPC platforms. Section 4 gives several examples and case studies showing how large-scale distributed platforms based on the grid and cloud computing paradigms have been used successfully in CDD projects. Section 5 presents our conclusions and discusses the future perspectives for drug discovery in combination with advanced computational techniques.

## 2. Virtual screening methods

This section reviews some of the methods – such as QSAR, docking and molecular docking, and MD – that are run over distributed computing infrastructures and used routinely in VS calculations.

### 2.1. QSAR

QSAR models can be defined as regression or classification models that relate several variables, called descriptors, with bioactivity values to predict the activity of new compounds. These descriptors codify several chemical features of compounds, including their physicochemical properties and experimental measurements. QSAR models are developed using different computational strategies [12] – such as statistical methods, which include multivariate linear regression analysis (MLR) [13], principal components analysis (PCA) [14], partial least-squares (PLS) analysis [15], and linear discriminant analysis (LDA) [16] – or artificial intelligence approaches, such as extreme learning machines (ELMs) [17], neural networks (NN) [18], and support vector machines (SVM) [19].

The applicability of QSAR models in drug discovery requires that the QSAR model has good predictability and provides a physicochemical interpretation of the possible mechanism of action. The development of QSAR models, their application to large datasets, the calculation of quantum chemical descriptors, and the use of advanced statistical methods or artificial intelligence approaches demands large amounts of computing resources [20] and mandatory access to HPC resources.

Tetko et al. [21] reviewed various public modeling environments for the development of QSAR models, such as OpenMolGRID [22,23] and its successor Chemomomentum [24] for grid computing, and their application to aquatic toxicity [25], acute toxicity [26], and the discovery of HIV-1 protease inhibitors [27]. Other QSAR tools running on distributed computing environments include: Simplex Representation of Molecular Structure–Structural and PhysiCochemical Interpretation [28], which can exploit ensemble predictions, classification techniques, and nonlinear methods, although the model-building parameters are kept fixed; ISIDA/QSPR [29,30], which also exploits ensemble predictions; and DTC Lab. Software Tools [31], which includes a validation method for the selection of models, is also valid for nonlinear techniques.

## 2.2. Molecular docking

Ligand-based methods (e.g. QSAR, similarity searching, pharmacophore modeling and docking) represent worthwhile solutions in drug discovery. However, QSAR and similarity searching do not take into account knowledge about the binding site within the protein target and this can reduce the accuracy of the calculations. To overcome this issue, structure-based methods are the preferred choice when the 3D structure of the target is known, although they are usually computationally more expensive than ligand-based approaches. In such cases, it is studied how the activity of proteins may be altered when small ligands dock into the well-defined cavities of protein receptors. These ligands can act as molecular switches and control the activity of the protein. For proteins involved in a metabolic pathway related to a disease, artificial ligands can act as drugs [32]. As more metabolic pathways and their associated key proteins are identified, the search for artificial ligands has intensified as a method of improving the treatment of various diseases. The number of known protein structures continues to grow exponentially, a trend increasingly complemented by initiatives in structural genomics [33]. Molecular docking identifies the lead compounds that can bind to a target protein with high affinity [34]. This is achieved by calculating the optimum binding position for each molecule in a large database of potential targets using heuristics and then ranking the database with a scoring function according to the estimated affinity [35].

Docking methods have been investigated for many years, and several compounds have been identified and developed as drugs [36]. Several docking methods are currently available – including AutoDock4 [4], AutoDock Vina [37], Glide [38] and Lead Finder [39] – each of which has different scoring functions and optimization methods. All of these methods use an atomistic representation of the protein and the ligand and allow the exploration of thousands of possible binding positions and ligand conformations in the coupling process [40,41]. As a result, the binding modes for many complexes are reliably predicted. However, unbiased comparative evaluations of the estimations of affinity by molecular docking show little correlation between the measured and predicted affinities over a wide range of receptor–ligand complexes [42], suggesting that more advanced approaches are required to increase the accuracy of the total binding energies. The use of explicit solvent molecules and the addition of dynamic effects to the system may partially circumvent the limitations associated with classical docking simulations.

## 2.3. Molecular dynamics

The dynamic nature of real biological environments needs to be considered if meaningful predictions are to be made. As it is not possible to apply the most demanding simulations directly to large libraries of compounds, an efficient workflow should start by first using inexpensive techniques, such as QSAR, and then proceed with more expensive molecular docking techniques. The resulting best-ranked drug candidates can then be selected and implemented in more advanced simulations, such as MD. This approach can include the combined effects of the solvent and temperature in the evolution of the system over relatively long-

time trajectories – that is, on the scale of nanoseconds to microseconds. MD has been used to rank a series of biologically active ligands docked into the herceptin antibody, an efficient biological molecule able to localize malignant cells in patients with breast cancer [43]. The MD simulations allowed ligands that produced an early release from the binding site (during first 100 ns) to be discarded because they were incompatible with a stable interaction.

MD has both advantages and disadvantages in drug discovery research. It has the advantage that the stability of binding sites can be validated before cell-based assays are carried out, but the production of MD trajectories requires large amounts of computational resources, usually defined as the number of nanoseconds computed per day. Huge efforts have been devoted to speed up simulations and most of the currently available MD codes include GPU-accelerated versions. Unfortunately, such codes require previous expertise from the user because molecules with chemical features that are not implemented in the force field parameters need to be optimized/implemented by the user. As an alternative, some web servers running under distributed computing infrastructures are available that may help non-expert users to perform short MD simulations in a friendly framework, which might be used as a first proof of concept. A series of representative examples of online solutions are MDWeb [44], the Gromacs server as implemented in Haddock [45], Vienna-PTM [46], CABS-flex [47], Protein structure REFinement via MD [48], MoMA-LigPath [49], and MoSGrid [50]. Most of these solutions offer predefined protocols to guide the preparation of the structure (i.e. the experimental PDB file), which should be ‘cured’ and then transformed to a specific format to run standard MD simulations. These listed tools also allow the trajectories produced to be analyzed by monitoring the stability of the whole system using the root-mean-square deviation (RMSD) as well as the geometrical parameters of the binding site (the non-covalent interactions between the target protein and ligand).

One of the major drawbacks of classical MD calculations is the assessment of entropic terms, thermodynamic parameters that are required to determine reliable absolute free energies. There is no universal solution for the refinement of the MD protocol, and several methods have been proposed to better capture the ligand–protein problem in drug discovery, including free energy perturbations, umbrella sampling, the potential of the mean force, and metadynamics. All these additional descriptors may be used to produce a more accurate prediction. However, their computational cost restricts their application to small and rigid systems, which, in turn, prevents the implementation of MD techniques in the servers currently available. Further development is needed before MD can be systematically applied to complex systems because it requires sampling at large dimensions to produce accurate values for the entropic and solvation contributions to the free energy [51].

## 3. Distributed computing environments

Biomedicine was one of the first areas of research [52] to move from the use of in-house computing facilities or single supercomputing centers to distributed infrastructures, in particular grid and

cloud platforms. These infrastructures are based on the vision of providing services to users through the sharing of capabilities and resources. The core idea is that any simulation can be achieved using the concept of service: a workstation or a supercomputer represents a computing service, but a database, a domain-specific application or authentication mechanisms are also services. It is possible to find dozens of published definitions of these platforms, which have been modified during their development. Foster [53] – one of the first developers of grid computing technologies – defined a computational grid as an infrastructure of both hardware and software that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities. Foster [53] suggested that the essence of grid computing can be described by three concepts: (1) the coordination of resources that are not subject to centralized control (see Figure 1); (2) access to, and the management of, resources using standard, open, general purpose protocols and interfaces; and (3) the delivery of non-trivial qualities of service. Many years after the spread of this technology began, we can now describe grid computing as a wrapper with which to freely access remote multi-institutional resources, with either dedicated or shared computational time. Grid computing paved the way for cloud computing.

Cloud computing has been defined by the US National Institute of Standards and Technology as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or interaction with the service provider. It is based on three service models (see Figure 2): (1) the software as a service (e.g. Dropbox); (2) the platform as a service (e.g. the Google App Engine); and (3) the infrastructure as a service (i.e. the possibility of managing virtual machines). In contrast with the definition of grid computing, cloud computing can be defined as a way to use the internet to deliver on-demand, scalable, pay-per-use services that rely on computational facilities hosted by a single institution.

Cloud infrastructures can be used for business, commercial, or research purposes, but the concept is always based on access to on-demand resources on a pay-per-use basis. By contrast, grid platforms are used for research by virtual organizations [54], which are dynamic sets of individuals and/or institutions sharing a goal to be pursued using the grid resources without any additional charge. The users of a cloud infrastructure are customers, but the users of a grid platform are members of one or more virtual organizations.

The following sections analyze the relevant tools and infrastructures based on the use of non-local computational resources, with a particular focus on architectures that have been used for CDD projects.

### 3.1. Grid computing

Although initially exploited in the field of high-energy physics, the biomedical community showed an immediate interest in grid computing – for drug discovery applications in particular [55] – and many computational challenges have been met in this context. In addition to large campaigns for the analysis of biosequences [56], grid computing has been widely exploited in structural biology for large-scale VSs of neglected diseases [57,58] and to determine the MD of huge biomolecular systems [59].

#### 3.1.1. Service-based grid computing

Most of the grid infrastructures rely on middleware toolkits – that is, a software layer that lies between the operating system and the applications. The first example of middleware was the Globus Toolkit [60], followed by others such as the Advanced Resource Connector and the Uniform Interface to Computing Resources. Many different platforms have been proposed on top of these middleware components for drug discovery projects and these usually follow one of two approaches.

The first approach relies on general purpose grid service environments coordinated by virtual organizations. This

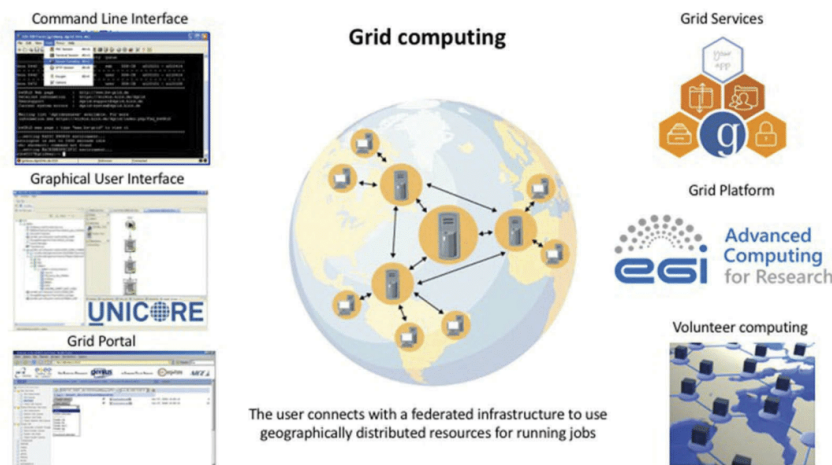
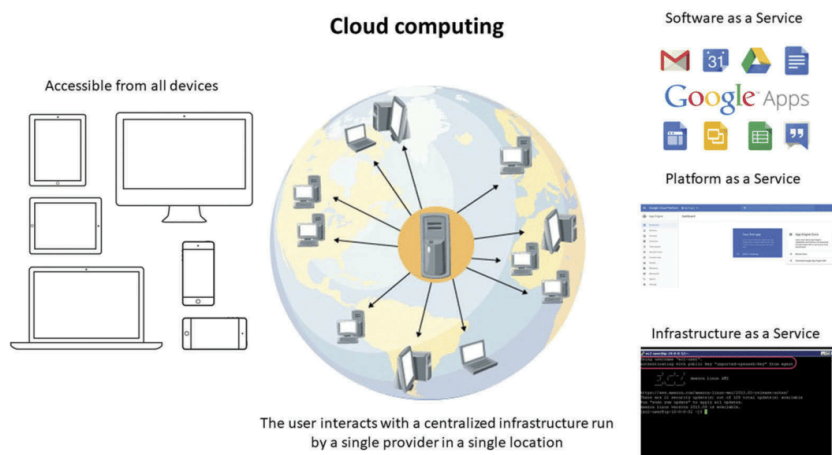


Figure 1. The Grid computing paradigm. On the left side, three examples of user interface are shown (Command line interface, graphical user interface, grid portal). The central panel shows the common federated and geographical dispersed infrastructure typical of grid computing. On the right side, the three general types of grid infrastructures are reported (Grid services, grid platform, volunteer computing).



**Figure 2.** The Cloud computing paradigm. On the left side, the figure shows how cloud infrastructures are easily accessible from any kind of device. The central panel shows the common centralized and single-entry point infrastructure typical of cloud computing. On the right side, three popular examples of cloud platforms are shown (Software as a service, platform as a service, infrastructure as a service).

paradigm can be exploited through both service and software frameworks, such as the distributed infrastructure with remote agent control (DIRAC) [61], the distributed analysis environment (DIANE) [62], and the gLite workload management system (WMS) [63], which are able to provide high-level job submission services at the front end and the possibility of interacting with heterogeneous systems at the back end (see Figure 1). The second approach aims to standardize the creation, representation, and sharing of the so-called computational workflows that tie a number of software tools together into a single analysis. This is built on application-oriented grid service infrastructures where the focus is on the software executed over the clusters provided within a grid (see Figure 1). This is the case for myGrid [64], a widely used infrastructure for composing pipelines in the field of bioinformatics and structural biology using the Taverna Workbench or the Virtual Laboratory for e-Science (VL-e) [65], which rely on a grid service infrastructure and use Nimrod-G semantics to compose complex workflows.

Service-based grid computing is usually free, but the user needs to belong to an academic institution. Association with a virtual organization is required for grids relying on such organizations.

### 3.1.2. Desktop-based grid computing

The other grid approach that gained progressive success in the scientific community is known as desktop grid or volunteer computing because it often relies on the general public to donate resources. Desktop computing and, more generally, volunteer computing are usually free and only registration is required (see Figure 1).

This paradigm relies on middleware oriented to peer-to-peer architectures. This means that there is no clear distinction between the grid entities owned by users (e.g. a workstation) to access the infrastructure (e.g. a cluster shared via grid middleware) and every resource can act as both a client and

a server at any one time. Any user can bring resources into the grid because the installation and maintenance of the software is intuitive, requiring no special expertise, which enables a large number of donors to contribute to the pool of shared resources.

The most popular middleware for desktop grid applications is the Berkeley Open Infrastructure for Network Computing (BOINC) [57], which was originally developed in the context of the SETI@home project [66] to search for extraterrestrial intelligence. The main feature of BOINC is represented by its simple API, which allows easy interaction with other environments, and its great community support. For example, the EDGeS project [67], which aims to create an infrastructure combining the advantages of service (the EGEE infrastructure) and desktop grids, is using BOINC.

### 3.2. Cloud computing

Grid-computing infrastructures are not completely satisfactory when running complex applications, in particular for industrial companies [68], because they do not provide a flexible and reliable environment. Most of the users submit batch jobs to remote clusters with little or no interactivity or the possibility of customizing the environment. The management of the storage of distributed data is complex, especially the administration of geographically dispersed databases. Misconfigured nodes on the grid are common, which results in jobs that fail continuously, emptying the grid queue, an effect known as shrink hole.

By contrast, cloud computing aims to provide the elements of a classical computational infrastructure (from a single workstation running the company website to a fully operational HPC cluster with high-speed network connections) as an on-demand service using virtual machines and virtual clusters, with development and execution frameworks and applications



as services (see Figure 2). This paradigm addresses the key demands of creating an easily accessible and flexible environment (see Figure 2), able to support data processing and, in general, to provide everything required to perform complex analyses [69]. Cloud computing vendors such as Amazon [70] and Google [71] provide specialized environments to ease biomolecular data processing in the cloud.

Cloud computing allows the adoption of novel programming models – such as MapReduce and Spark, both of which rely on the Hadoop file system, an open source framework that enables the distributed processing of large datasets – which are exploited for sequence alignment [72–74], drug discovery [75–77], and many other applications [78]. In MapReduce, the input of a computation is split into independent chunks, which are then processed by the map tasks in a parallel manner. The results are then sorted and processed by the reduce tasks to provide the final output. By contrast, Spark works as a toolset for distributed programs and offers a (deliberately) restricted form of distributed shared memory, facilitating the implementation of iterative algorithms that visit their dataset multiple times.

Some initiatives support the execution of virtual machines on top of grid resources, such as the Worker Nodes on Demand Services (WNoDes) framework [79]. This approach has been exploited for macromolecular characterization and the estimation of free energy in protein–protein docking [80].

### 3.3. GPUs and distributed computing

GPUs can be used to reduce the execution time of scientific applications [81,82] within distributed computing scenarios. In particular, GPUs have been ranked as one of the platforms with the highest projection for implementing algorithms that simulate complex scientific problems. Since their first appearance, the development of GPUs has been marked by the world of video games, which have reached high levels of popularity as more realism is achieved. In 2006, NVIDIA, the leader in the manufacturer of GPUs, made a breakthrough in the world of HPC when they released the Compute Unified Device Architect (CUDA) development kit. This architecture makes it possible to use GPUs for the development of scientific applications. Six of the ten most powerful supercomputers in the world [83] currently have coprocessor (GPU or similar vector-based devices) accelerators.

In the limited subset of problems for which enough effort can be invested to ensure that specific drug discovery software supports GPU computing, these devices can greatly accelerate calculations. Not all algorithms are good candidates for acceleration [84,85], but scientists are now aware of the benefits that HPC can bring to their research, either by including these hardware platforms in their studies or by directly designing and implementing specific software that can be used on these platforms.

## 4. Examples of applications

This section discusses examples of drug discovery using distributed distributed computing architectures and ligand-based VS (LBVS) servers.

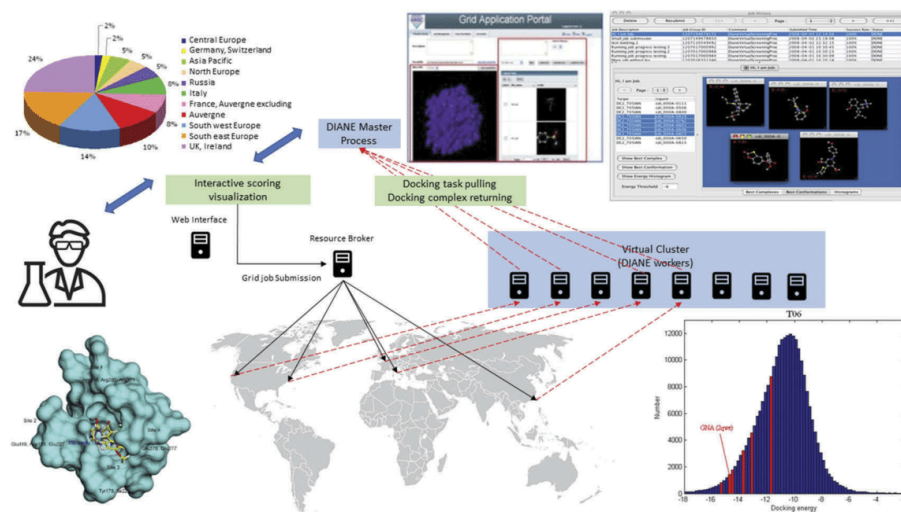
### 4.1. Grid-based examples

One of the most interesting grid projects is Wide in Silico Docking on Malaria (WISDOM) [86], which has the goal of using the computing resources offered by the Enabling Grids for E-Science in Europe (EGEE) project (the largest grid project in Europe developed for the Large Hadron Collider at CERN in Geneva) to select drug-like molecules active on a biological target to fight malaria [87]. Malaria was chosen because tropical diseases often suffer from a lack of research due to the cost of bringing new drugs to market. The project developed *in silico* approaches to VS, mainly using resources from the EGEE project, but also from EUCINAgrid and TWGrid in Asia, EUMEDGRID in Africa, and OSG and Digital Ribbon in the USA [58]. Several of the drug-like molecules selected *in silico* have been confirmed by *in vitro* tests to be active inhibitors and the most promising molecules have been patented [88].

An infrastructure similar to that used in WISDOM was also used to perform a large VS project (see Figure 3) aimed at identifying new drugs for potential variants of the influenza A virus [57]. About 300,000 compounds selected from the ZINC database and ad hoc chemical combinatorial library were screened against eight variants of neuraminidases predicted by homology modeling. The distributed analysis environment was used as the job dispatcher and a portal was developed to visualize the achieved outcome to make the docking conformations and scoring results available to biologists. Using thousands of CPU cores on a grid, a six-week long, high-throughput screening activity was accomplished, performing >100 CPU core years of calculations, producing around 600 gigabytes of results, and identifying about 100 compounds for further biological analysis and testing (see Figure 3).

Among the grid initiatives oriented toward structural biology analysis and relevant to drug discovery, the European Model for Bioinformatics Research and Community Education (Embrace) network of excellence [89] has been particularly important in the promotion, development, and implementation of standards for the interoperability of resources in the life sciences community. Another example is the BIOINFOGRID project [90], which aims to exploit the existing EGEE infrastructure for large-scale modeling and simulations of biological problems and in which docking and MD had a primary role. Many attempts have been made to perform MD on grid [91,92] and cloud [93,94] platforms, although the only approach that gained some popularity was the Hamiltonian Replica Exchange method [95].

In addition to dedicated virtual organizations, several ad hoc grid systems have been proposed for the execution of bioinformatics and drug discovery workflows [96]. One of the most effective service-based environments for running workflows is Taverna [97]. Many projects oriented toward the annotation, modeling, and simulations of biological macromolecules, such as myGrid [64] and the VL-e [65], rely on this infrastructure. Taverna was designed to combine distributed grid services and/or local tools into complex analysis pipelines to tackle a wide range of scientific research [98]. Once constructed, the workflows become reusable resources. Many of these workflows are oriented toward structural biology and



**Figure 3.** The architecture of the grid-based infrastructure adopted for WISDOM and the Avian flu virtual screenings. DIANE was used as job dispatcher and a grid application portal was developed to visually inspect achieved results. In the top left corner, the distribution of the jobs among the different geographical regions of the world. In the bottom left corner, the structure of the neuraminidase that was target of the screening. In the top right corner, the job monitoring application used in the virtual screening. In the bottom right corner, the enrichment curve of one of the leading compounds identified with the virtual screening.

drug discovery, in particular the VS and identification of protein-binding sites, by leveraging Taverna's technology for data storage, workflow enactment, change event notification, resource discovery, and provenance management [99].

The VL-e environment, by leveraging existing grid technologies, enables molecular modeling for drug design using geographically distributed resources. The concept involves screening millions of compounds in a chemical database against a protein target to identify those molecules with potential use in drug design. The VL-e uses the Nimrod-G parameter specification language to transform the existing molecular docking application into a parameter sweep application suitable for execution on distributed systems. New tools have been developed to enable access to ligand records/molecules in chemical databases from remote resources. The Nimrod-G resource broker, along with the chemical database data broker, is used for scheduling and the on-demand processing of docking jobs on the grid resources. The results show the ease of use and power of the Nimrod-G language and VL-e tools for drug discovery on grid-computing platforms [100].

Moving toward the desktop grid paradigm, one important example is represented by the Drug Discovery Grid (DDGrid) [101] project. This focuses on providing drug screening services, such as building docking processes for the virtual high-throughput screening of the avian influenza virus [102]. The project relies on a grid environment and the grid computation resources of the Grid@Asia project. DDGrid leverages BOINC with grid technologies to harness the power of clusters and super-computing systems owned by different organizations in China and South Korea.

Rosetta@home, a distributed computing project for the prediction of protein structures, also relies on the BOINC platform and aims to model protein-protein docking and protein

structures. Rosetta@home has been used to validate proteins that neutralize influenza [103,104] and enzymes that breakdown gluten for the treatment of celiac disease, all of which are moving through animal trials and into clinical trials [105,106]. Rosetta@home has also been applied to research on malaria, Alzheimer's disease, and other pathologies [107].

In addition to disease-related research, the Rosetta@home network serves as a testing framework for new methods in structural biology. Such methods are then used in other Rosetta-based applications, such as RosettaDock [108], RosettaDesign, and Robetta [109]. Rosetta@home consistently ranks as one of the foremost docking predictors and is one of the best predictors of tertiary structure currently available [110]. Using hundreds of thousands of cellular telephones and other mobile devices, Rosetta@home was used to validate peptide macrocycles designed with rigid structures, which could be useful as peptide therapeutics [111,112].

The conformational states from Rosetta's software can be used to initialize a Markov state model as starting points for simulations using Folding@home [113], which is a distributed computing platform for studying protein folding and other types of problems that can be solved with MD. As Rosetta only tries to predict the final folded state, and not how folding proceeds, Rosetta@home and Folding@home are complementary and address very different molecular questions. Folding@home has been recognized as the most powerful distributed computing network in the world [114] and, in contrast to other similar projects, it does not rely on BOINC, but on a specific networking library called Cosm [115].

Large-scale executions of drug discovery applications on volunteer grids are no longer simply research projects, but an effective way to run *in silico* simulations. The main issue is

the ratio of volunteer resources needed to achieve the computing power of a grid or cloud node, which is, on average, greater than three [116,117].

#### 4.2. Cloud-based examples

Many examples of cloud computing rely on the platform as a service (PaaS) model, where applications are built using higher level platforms/frameworks. Many MapReduce-inspired frameworks developed by the Apache Software Foundation are currently available to manage and process large amounts of high-throughput omic data. For example, the Cancer Genome Atlas project made use of Hadoop to split genome data into chunks distributed over the cluster for parallel processing [118,119].

The Collaborative Genomic Data Model (CGDM) [120] uses Hadoop to boost the querying speed for the main classes of query on genomic databases. MetaSparks [120] uses Spark to recruit large-scale metagenomics reads to reference genomes, achieving better scalability and sensitivity than programs based on a single machine [121], a principle that has also been applied to drug discovery [122].

#### 4.3. GPU-based examples

GPUs have recently started to be used outside their traditional role as a graphical component of computers. This new application is known as general purpose computing on graphics processing units (GPGPU).

An example of GPU sharing within a distributed infrastructure is GPUGRID.net [123], a distributed computing infrastructure devoted to biomedical research. It consists of many GPUs joined together to deliver high performance all-atom biomolecular simulations. The simulations aim to develop better drugs by determining the mechanisms of drug resistance in cancers, modeling HIV maturation and investigating the features of neurologically important proteins.

The possibility of accessing GPUs on distributed facilities is particularly appealing because the most popular manufacturer, NVIDIA, markets itself as the best way to accelerate this kind of application. They have reported that MD applications, such as ACEMD [124] and GROMACS [125], traditionally CPU-based, achieve increases in speed of up to eight times with the incorporation of their massive parallel platforms.

Researchers have carried out comparisons showing the reduced times achieved by performing tasks in these new platforms relative to traditional computation. Marin et al. [126] compared the performance of a parallelization on multi-core systems against a GPU and showed that the GPU implementation was up to 33 times faster. Analyses by Chiappori et al. [93] showed that these simulation techniques are time-consuming and, therefore, parallel computing and GPU computations are required to reduce computation times.

Interesting papers have been published in this area of research – for example, Hung et al. [127] reviewed the two main CADD-based approaches (structure- and ligand-based drug design) and concluded that both multiple computers and GPGPU approaches can significantly improve the CADD

performance. Vogt et al. [128] reviewed the current trends in method development, including the implementation of GPUs.

Ma et al. [129] used GPUs to accelerate the chemical similarity calculation that plays a major part in VS. Malhat et al. [130] saved between 76 and 99% of their computation time by using GPUs in their implementation of the Ward algorithm to group similar chemical compounds. Lo et al. [131] used CUDA to accelerate the prediction of protein–ligand-binding regions using geometrical features.

Docking applications based on the scoring function, such as MetaDock [132], use a metaheuristic scheme to generate a large number of heuristic strategies for VS. MetaDock is designed to take full advantage of parallel and heterogeneous architectures, including multiprocessor chipsets and NVIDIA GPUs. Figure 4 shows how the MetaDock data distribution model works in GPUs. In this schema, the receptor molecule is stored in the shared memory of the multiprocessors to make better use of the memory accesses. The drug candidates are grouped in blocks of threads and are able to share the data from the receptor between the threads of the same multiprocessor. The application uses a computational molecular coupling methodology that seeks to predict the non-covalent binding of molecules or, more often, a macromolecule (receptor) and a small molecule (ligand). The aim is to predict the bound conformations and affinity of the union – that is, the strength of association – which is usually measured by a scoring function [133,134].

Another approach that takes advantage of parallel programming is the migration of tools into this new parallel paradigm. McIntosh-Smith et al. [135] used this new GPU programming paradigm to increase the overall performance of the drug-screening process, porting their tools to OpenCL. They describe the BUDE (Bristol University Docking Engine), which is a structure-based VS (SBVS) engine, and present it as one of the first applications migrated to modern hardware [135]. Fang et al. [136] describe GeauxDock, a new molecular docking program migrated to run over parallel platforms. Some of these research groups are also involved in the review of new programs developed to take advantage of these parallel platforms [137]. Krige et al. [138] ported a commercial application (blazeV10) of VS to these platforms and compared an OpenCL version on a broad range of devices. Harvey et al. [139] implemented ACEMD, a production-class biomolecular dynamics simulation program specifically designed for GPUs. Sukhwani et al. [140] accelerated a production mapping software program using NVIDIA GPUs.

In addition to migrating tools to a distributed infrastructure to take advantage of the massive parallel platforms, another approach is to implement new programs to perform some tasks. For example, McArt et al. [141] implemented a new program to perform connectivity mapping, which is a computational technique dedicated to drug repurposing around differential gene expression analysis. They drastically reduced the computational times; previous implementations took up to seven days, whereas using a GPU reduced the time required to just 10 min.

Studies in parallel algorithms that have led to new advances in this area include the work of Hu et al. [142], who developed a new GPU-based analytical method focusing on risk epistasis in a genome-wide association study of complex traits.

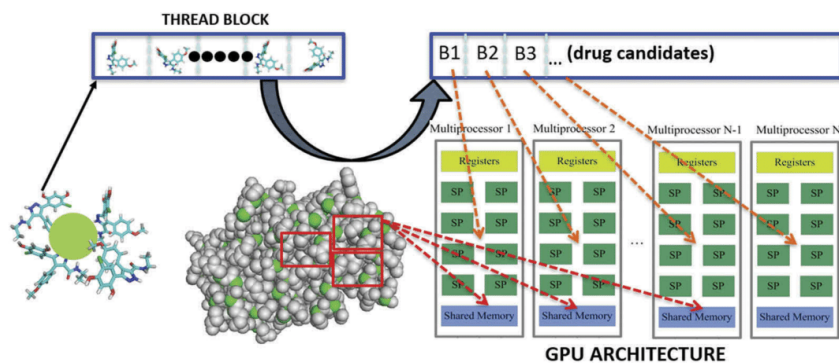


Figure 4. Data distribution model on GPU in METADOCK. Drug candidates are grouped in blocks of threads and mapped to the multiprocessors for their computation. The shared memory, in the lower part, stores parts of the receptor molecule for reuse by the candidates that integrate the same block of threads.

#### 4.4. Applications of virtual screening web servers

HPC has been extensively applied in VS. When full information about the target protein is available, SBVS methods are preferred and several servers are available to perform docking calculations. DOCK Blaster [143], Haddock [45], iScreen [144], SwissDock [145], and VSDocker [146] are examples of this type of server. Achilles [147] is a web tool implementing the blind docking approach. A distinguishing feature of Achilles is that it provides detailed reports, including the most likely ligand structure clusters, after processing the whole surface of the chosen protein.

SBVS calculations are usually very accurate, but the required information from the target protein is not always available. In such cases, LBVS is used as an alternative approach. A number of different servers implementing different approaches to LBVS are available on the web. As an example, SwissSimilarity [148] performs both 2D and 3D similarity searching, whereas HybridSim-VS [149] and USR-VS [150] focus on 3D similarity, including geometrical information about molecules. Other servers, such as iDrug [151], ChemMapper [152] and ZINCPharmer [153], use pharmacophore modeling to assess similarity. LBVS servers can efficiently scan millions or billions of compounds in a short time and are often a cost-effective solution to screening.

These examples have focused on the core of VS calculations. However, HPC can also help in the early stages of the screening process. For example, the prediction of molecular descriptors or fingerprints (e.g. molecular weight, number of rotatable bonds, number of acceptor/donor hydrogens) is a sensitive task that benefits from high computing power. BioTriangle [154] is a complete toolkit for the characterization of complex biological molecules and their interactions. This tool is not only useful in the field of cheminformatics, but also in bioinformatics. ChemDes is another web tool for fingerprinting and molecular descriptor calculations. It hides the complexity of many other open source packages from the users. An extensive compilation of tools and databases for drug discovery is maintained by the Swiss Institute of Bioinformatics [155].

#### 5. Conclusions

Distributed computing infrastructures, in particular those relying on GPUs, are very important in drug discovery programs. Although Big Pharma can rely on proprietary infrastructures to perform their analyses in-house, small- or medium-sized biotechnology laboratories need to exploit distributed infrastructures to achieve a flexible and cost-effective platform to perform their simulations. With distributed computing platforms, all players can combine the use of third-party software made available through a web interface with their own tools to exploit the full potential of this approach. There is a wide range of online VS servers providing services for many of the steps in the drug discovery process, including screening and fingerprint calculations.

The use of GPUs as computing platforms for drug discovery processes will further accelerate computations, as shown by the fact that commercial tools now include GPUs in their implementation to take advantage of their enhanced performance. Working with heterogeneous computation resources can help to improve the performance of applications. The main issues in these technological approaches are related to the cost-effective exploitation of the available computing capabilities because each GPU model has different features depending on its family and generation and variations in the number of cores and performance. There is a need to develop applications with a load-balancing technique to correctly account for the characteristics of each device when distributing the workload through a distributed computing environment. Grid and cloud science gateways are an effective solution to providing user-friendly access to computational power and tools, while solving security issues and the problem of moving data between centers. These infrastructures are typically free for scientific purposes.

#### 6. Expert opinion

Rapid developments in computational capabilities, combined with the explosion of research into personalized medicine, are

currently leading the third wave of drug discovery. This new approach of personalized drug discovery should greatly improve the therapeutic effects of drugs while reducing their side-effects. The screening and validation of functional gene and protein targets in the early stages of drug discovery, and the development and optimization of selected molecules, requires huge amounts of computational power, which is difficult to buy and operate for small- and medium-sized biotechnology laboratories and academic research groups. In addition, perhaps surprisingly, large pharmaceutical companies are increasingly outsourcing research and computing activities to cut costs and to access state-of-the-art knowledge and technologies. We, therefore, see leading scientific institutions and computational chemistry companies shifting their business model from releasing commercial packages to providing distributed/cloud support for drug discovery research, while pharmaceutical companies exploit cloud services for drug development. As an example, in 2014, the Novartis Institutes for Biomedical Research used Amazon infrastructures to build a platform for the VS of compounds against a common cancer target [156], leveraging the power and agility of cloud computing to conduct 39 years of computing in just 11 hours, creating an on-demand computing system that would cost an estimated US\$44 million to build for a cost of only US \$5000. It is now almost impossible for researchers to enter the world of drug discovery without using these platforms.

These developments lead to the key message of this paper: recent advances in distributed computing technologies and infrastructures mean that they are now essential in the field of drug discovery. This was not true before the emergence of the grid computing paradigm about 30 years ago and, later, the cloud-computing paradigm, but is currently very clear. The possibility of exploiting these paradigms presents great opportunities, as demonstrated by the large simulations conducted on grids and clouds by both pharmaceutical companies and public institutions, but also presents some weaknesses in terms of the effectiveness, accuracy, and usability of software and hardware in distributed environments.

Effectiveness and accuracy mean the use of the correct methods in QSAR, VS, lead optimization, and MD. Therefore, after the development of new algorithms, usually by academic researchers, it is necessary to support the software through its full development for commercial use. The most common business model is to create spin-off companies to consolidate the development of these approaches, closing the gap between academia and the market, with sponsorship from pharmaceutical companies. These companies should also support users in applying these new approaches in the correct way, preventing failures in handling each specific tool. Some successful examples are available in the field of drug discovery, such as AMBER, one of the most popular software packages for MD. Usability refers to the development of environments that lower the barriers to applying distributed and parallel computing infrastructures for drug discovery and offer support in handling failures.

Despite some open issues, drug discovery is moving toward distributed infrastructures because they provide high performance and cost-effective access to software tools, data, and infrastructures. From the computational point of view, *in silico* analyses that were, until recently, impossible, are becoming increasingly feasible by applying the most suitable distributed

computing infrastructure. This represents an incredible potential for this field of research, but more effort is required to develop and automate the functionalities that are crucial in enabling agile and flexible predictive modeling and simulation protocols. Workflow management systems can aid in many of these challenges, but the currently available systems are not suitable for users unfamiliar with ICT systems.

By considering the cost-effectiveness of the different distributed infrastructures available to researchers for drug discovery, a recent review of computational models in computational biology evaluated the performance of a single application exploiting grid and cloud computing [69]. This study reported comparable results in terms of execution time, but also demonstrated that grid platforms have overheads due to failures caused by server misconfigurations and waiting times in cluster queues. Cloud infrastructures should, therefore, be preferred over grid approaches if the budget is sufficient, whereas if cost is a major issue, such as for neglected diseases, grid platforms may still be a valid solution.

Cloud platforms have cost advantages compared with buying local facilities because dynamic access to resources is less expensive than having a private infrastructure. In particular, if spot instances – cheap resources that can be turned off by the vendor when there is a high on-demand computational load – can be used, which may occur in VS but not in MD simulations, then cloud computing could be cheaper than buying computing power from a high-performance computational facility. This is particularly true if the cloud servers are equipped with high-end GPU devices and the applied cost model is appropriate.

Grid infrastructures and supercomputing centers require the pre-installation and tuning of software, whereas cloud solutions can exploit the pre-configured virtual machines released by vendors (e.g. Amazon and Google) or specialized companies (e.g. Eagle Genomics or Cloud Pharmaceuticals). However, it is important to highlight the fact that containers specialized for drug discovery analysis – for example, those relying on Docker [157] – can be used to rapidly create custom environments on all these platforms, reducing the time needed to set-up and configure the software [158].

Future work in this area should be directed toward simplifying analyses and improving cooperation among researchers by providing input data, sharing the parameterization settings, and releasing non-sensitive results to the scientific community. Open access to data and the reproducibility of methods have gained much attention in recent years and this is particularly true for approaches to developing new drugs. Tools are being developed to support researchers in this regard, but they require improvements and the trust of the scientific community. Workflow management systems and virtualization platforms are currently available to help improve the reproducibility and sharing of results and methods, although this is seldom considered before a publication or a patent is accepted as a result of the costly and long-running analyses required to achieve significant results and due to privacy/security issues. We postulate that open sharing is beneficial to the research community and research is more efficient when sharing is possible while protecting sensitive data.

A particular area of interest that, in our opinion, represents the next revolution in drug discovery is related to deep-

learning approaches [159], both for selecting compounds and for their optimization. With the support of NVIDIA, some tools have already been released to improve the quantum mechanics energy function and to produce computationally fast and accurate molecular energy surfaces, geometries, and forces. As a result of the complexity of these approaches, the use of HPC in a distributed infrastructure will be unavoidable.

### Funding

This work was funded by a grant from the Spanish Ministry of Economy and Competitiveness [CTQ2017-87974-R] and by the Spanish MEC and European Commission FEDER [TIN2016-78799-P, AEI/FEDER, UE]. This research was partially supported by the supercomputing infrastructure of Poznan Supercomputing Center, the e-infrastructure program of the Research Council of Norway, the supercomputer center of UiT – the Arctic University of Norway and by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA–CIEMAT), funded by the European Regional Development Fund (ERDF). CETA–CIEMAT belongs to CIEMAT and the Government of Spain. The authors also acknowledge the computing resources and technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga. Powered@NLHPC research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02).

### Declaration of interest

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

### Reviewer disclosures

One referee is an employee of Merck & Co.

### ORCID

Antonio Llanes Castro  <http://orcid.org/0000-0002-9802-4240>

### References

Papers of special note have been highlighted as either of interest (\*) or of considerable interest (\*\*) to readers.

- Naylor CB, Richards WG Computer-aided drug design [thesis]. Oxford (UK): University of Oxford; 1984.
- McCammon JA. Computer-aided molecular design. *Science*. 1987;238(4826):486–491.
- Shoichet BK. Virtual screening of chemical libraries. *Nature*. 2004;432(7019):862–865.
- Morris GM, Huey R, Lindstrom W, et al. AutoDock4 and autoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem*. 2009;30(16):2785–2791.
- Friesner RA, Murphy RB, Repasky MP, et al. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*. 2006;49(21):6177–6196.
- Merelli I, Cozzi P, D'Agostino D, et al. Image-based surface matching algorithm oriented to structural biology. *IEEE/ACM Trans Comput Biol Bioinform*. 2011 Jul-Aug;8(4):1004–1016.
- Merelli I, Morra G, Milanese L. Evaluation of a grid based molecular dynamics approach for polypeptide simulations. *IEEE Trans Nanobioscience*. 2007;6(3):229–234.
- Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? *Drug Discov Today*. 2002;7(17):903–911.
- Rudin M, Weissleder R. Molecular imaging in drug discovery and development. *Nat Rev Drug Discov*. 2003;2(2):123–131.
- Wang L, Wu YJ, Deng YQ, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc*. 2015;137(7):2695–2703.
- D'Agostino D, Clematis A, Quarati A, et al. Cloud infrastructures for in silico drug discovery: economic and practical aspects. *Biomed Res Int*. 2013;2013:138012.
- Roy K, Kar S, Das RN. A primer on QSAR/QSPR modelling. In: Chapter 2, *Statistical Methods in QSAR/QSPR*. 1st ed. Cham: Springer International Publishing; 2015. p. 37.
- Pérez-Garrido A, Girón-Rodríguez F, Helguera AM, et al. Topological structural alerts modulations of mammalian cell mutagenicity for halogenated derivatives. *SAR QSAR Environ Res*. 2014;25(1):17–33.
- Yoo C, Shahlai M. The applications of PCA in QSAR studies: A case study on CCR5 antagonists. *Chem Biol Drug Des*. 2018;91(1):137–152.
- Luco JM, Ferretti FH. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J Chem Inf Comput Sci*. 1997;37(2):392–401.
- Pérez-Garrido A, Helguera AM, Rodríguez FG, et al. QSAR models to predict mutagenicity of acrylates, methacrylates and alpha, beta-unsaturated carbonyl compounds. *Dent Mater*. 2010;26(5):397–415.
- Pérez-Garrido A, Girón-Rodríguez F, Bueno-Crespo A, et al. Fuzzy clustering as rational partition method for QSAR. *Chemometr Intell Lab Syst*. 2017;166:1–6.
- Ghasemi F, Mehrdehnavi A, Pérez-Garrido A, et al. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. *Drug Discov Today*. 2018;S1459–6446(17):30476–30482.
- Darnag R, Mazouz ELM, Schmitzer A, et al. Support vector machines: development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives. *Eur J Med Chem*. 2010;45(4):1590–1597.
- Darvas F, Papp Á, Bágyi I, et al. OpenMolGRID, A GRID based system for solving large-scale drug design problems. In: *Dikaiakos MDeditor. Grid computing*. Berlin, Heidelberg: Springer; 2004. p. 69–76.
- Tetko IV, Maran U, Tropsha A. Public (Q)SAR services, integrated modeling environments, and repositories on the web: state of the art and perspectives for future development. *Mol Inf*. 2017;35:1600082.
- A review of public modeling environments for QSAR development. The authors explain the transition from the old-fashioned models to the new online approach.**
- Sild S, Maran U, Lomaka A, et al. Open computing grid for molecular science and engineering. *J Chem Inf Model*. 2006;46:953–959.
- Sild S, Maran U, Romberg M, et al. OpenMolGRID: using automated workflows in GRID computing environment. In: Sloot PMA, Hoekstra AG, Priol T, et al. editors. *Advances in grid computing - EGC 2005*. Berlin, Heidelberg: Springer; 2005. p. 464–473.
- Schuller B, Demuth B, Mix H, et al. Chemomomentum – UNICORE 6 based infrastructure for complex applications in science and technology. In: Bougé L, Forsell M, Träff JL, et al., editors. *Proceedings of the Euro-Par 2007 Workshops Parallel Processing*; 2007; Rennes, France: Springer-Verlag; 2008; p.82–93.
- Mazzatorta P, Smiesko M, Lo Piparo E, et al. QSAR model for predicting pesticide aquatic toxicity. *J Chem Inf Model*. 2005;45:1767–1774.
- Maran U, Sild S, Mazzatorta P, et al. Grid computing for the estimation of toxicity: acute toxicity on fathead minnow (*pimephales promelas*). In: Dubitzky W, Schuster A, Sloot PMA, et al. editors. *Distributed, high-performance and grid computing in computational biology*. Berlin, Heidelberg: Springer; 2007. p. 60–74.
- Maran U, Sild S, Kahn I, et al. Mining of the chemical information in GRID environment. *Future Gener Comput Syst*. 2007;23:76–83.
- Polishchuk P, Tinkov O, Khristova T, et al. Structural and Physico-Chemical Interpretation (SPCI) of QSAR models and its

20  A. J. BANEGAS-LUNA ET AL.

- comparison with matched molecular pair analysis. *J Chem Inf Model.* 2016;56(8):1455–1469.
29. Varnek A, Fourches D, Horvath D, et al. ISIDA – platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr Comput Aided Drug Des.* 2008;4:191–198.
30. Ruggiu F, Marcou G, Varnek A, et al. ISIDA property-labelled fragment descriptors. *Mol Inform.* 2010;29(12):855–868.
31. Ambure P, Aher RB, Gajewicz A, et al. NanoBRIDGES software: open access tools to perform QSAR and nano-QSAR modelling. *Chemom Intell Lab Syst.* 2015;147:1–13.
32. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br J Pharmacol.* 2007;152(1):9–20.
33. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45(D1):D353–361.
34. Ferreira L, Dos Santos RN, Oliva G, et al. Molecular docking and structure-based drug design strategies. *Molecules.* 2015;20:13384–13421.
35. Kitchen DB, Decornez H, Furr JR, et al. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov.* 2004;3(11):935–949.
36. Talele TT, Khedkar SA, Rigby AC. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr Top Med Chem.* 2010;10(1):127–141.
37. Trott O, Olson AJ. AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem.* 2010;31(2):455–461.
38. Friesner RA, Murphy RB, Repasky MP, et al. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *J Med Chem.* 2006;49(21):6177–6196.
39. Stroganov OV, Novikov FN, Stroylov VS, et al. Lead finder: an approach to improve accuracy of protein–ligand docking, binding energy estimation, and virtual screening. *J Chem Inf Model.* 2008;48(12):2371–2385.
40. Chiappori F, D'Ursi P, Merelli I, et al. In silico saturation mutagenesis and docking screening for the analysis of protein–ligand interaction: the endothelial protein C receptor case study. *BMC Bioinformatics.* 2009;10(12):S3.
41. D'Ursi P, Chiappori F, Merelli I, et al. Virtual screening pipeline and ligand modelling for H5N1 neuraminidase. *Biochem Biophys Res Commun.* 2009;383(4):445–449.
42. Bock J, Gough D. A new method to estimate ligand–receptor energetics. *Mol Cell Proteomics.* 2002;1:904.
43. Cerón-Carrasco JP, Pérez-Sánchez H, Zúñiga J, et al. Antibodies as carrier molecules: encapsulating anti-inflammatory drugs inside herceptine. *J Phys Chem B.* 2018;122(7):2064–2072.
44. Molecular dynamics on web. Available from: <http://mmb.irbbarcelona.org>
45. Gcp VZ, Rodrigues JPGLM, Trellet M, et al. The HADDOCK2.2 web-server: user-friendly integrative modeling of biomolecular complexes. *J Mol Biol.* 2015;428:720–725.
46. Margreitter C, Petrov D, Zagrovic B. Vienna-PTM webserver: a toolkit for MD simulations of protein post-translational modifications. *Nucleic Acids Res.* 2013;41:W422–W426.
47. CAB-S flex server. Laboratory of theory of biopolymers, faculty of chemistry. University of Warsaw. Available from: <http://biocomp.chem.uw.edu.pl/CABSflex>
48. Heo L, Feig M. PREFMD: a web server for protein structure refinement via molecular dynamics simulations. *Bioinformatics.* 2018;34(6):1063–1065.
49. Devaurs D, Bouard L, Vaisset M, et al. MoMA-LigPath: a web server to simulate protein–ligand unbinding. *Nucleic Acids Res.* 2013;41(W1):W297–302.
50. Krüger J, Grunzke S, Gesing S, et al. The MoSGrid science gateway – a complete solution for molecular simulations. *J Chem Theory Comput.* 2014;10(6):2232–2245.
51. Chen PC, Kuyucak S. Accurate determination of the binding free energy for KcsA-charybdoxin complex from the potential of mean force calculations with restraints. *Biophys J.* 2011;100(10):2466–2474.
52. Chiappori F, Merelli I, Milanese L, et al. Static and dynamic interactions between GALK enzyme and known inhibitors: guidelines to design new drugs for galactosemic patients. *Eur J Med Chem.* 2013;63:423–434.
53. Foster I, Kesselman C. The history of the grid. *Adv Parallel Comput.* 2011;20:3–30.
- **This article describes the current status and future perspective of grid computing.**
54. Foster I, Kesselman C, Nick JM, et al. Grid computing: making the global infrastructure a reality: wiley series. In: Chapter 8, The physiology of the grid. New York: Wiley and Sons; 2003. p. 217–249.
55. Chien A, Foster I, Goddette D. Grid technologies empowering drug discovery. *Drug Discov Today.* 2002;7(20 Suppl):S176–80.
56. Trombetti GA, Merelli I, Orro A, et al. BGBlast: a BLAST grid implementation with database self-updating and adaptive replication. *Stud Health Technol Inform.* 2007;126:23–30.
57. Anderson DP. Boinc: a system for public-resource computing and storage. In: Proceedings of the Fifth IEEE/ACM International Workshop on Grid Computing 2004; IEEE; 2004. p. 4–10. Pittsburgh, PA, USA.
- **This article describes the BOINC distributed infrastructure.**
58. Jacq N, Breton V, Chen HY, et al. Virtual screening on large scale grids. *Parallel Comput.* 2007;33(4):289–301.
59. Chiappori F, Pucciarelli S, Merelli I, et al. Structural thermal adaptation of  $\beta$ -tubulins from the Antarctic psychrophilic protozoan *Euplotes focardii*. *Proteins.* 2011;80(4):1154–1166.
60. Foster I, Kesselman C. The globus toolkit. In: The grid: blueprint for a new computing infrastructure. San Francisco: Morgan Kaufmann; 1999. p. 259–278.
61. Fifield T, Carmona A, Casajús A, et al. Integration of cloud, grid and local cluster resources with DIRAC. *J Phys.* 2011;331(6):062009.
62. Germain-Renaud C, Loomis C, Moscicki J, et al. Scheduling for responsive grids. *J Grid Comput.* 2008;6(1):15–27.
63. Cecchi M, Capannini F, Dorigo A, et al. The gLite workload management system. In: International Conference on Grid and Pervasive Computing. Springer Berlin Heidelberg; 2009. P. 256–268. Geneva, Switzerland.
64. Stevens RD, Robinson AJ, Goble CA. myGrid: personalised bioinformatics on the information grid. *Bioinformatics.* 2003;19(Suppl. 1):i302–4.
65. Rauwerda H, Roos M, Hertzberger BO, et al. The promise of a virtual lab in drug discovery. *Drug Discov Today.* 2006;11(5–6):228–236.
66. Anderson DP, Cobb J, Korpela E, et al. SETI@home: an experiment in public-resource computing. *Commun ACM.* 2002;45(11):56–61.
67. Fedak G, He H, Lodygensky O, et al. EDGeS: a bridge between desktop grids and service grids. In: ChinaGrid Annual Conference, 2008. ChinaGrid'08; 2008 Aug 3; IEEE; 2008. p. 3–9. Dunhuang, China.
68. Merelli I, Cozzi P, Ronchieri E, et al. Porting bioinformatics applications from grid to cloud: a macromolecular surface analysis application case study. *Int J High Perform Comput Appl.* 2017;31(3):182–195.
69. Kasson PM. Computational biology in the cloud: methods and new insights from computing at scale. In: Pacific Symposium on Biocomputing, 2013. Kohala Coast, Hawaii, USA: World Scientific; 2012. p. 451–453.
70. Murty J. Programming amazon web services: S3, EC2, SQS, FPS, and simpleDB. Sebastopol(CA): O'Reilly Media, Inc; 2008.
71. Gruber K. Google for genomes. *Nature Biotechnology.* 2014;32:508.
72. Matsunaga A, Tsugawa M, Fortes J. Cloudblast: combining mapreduce and virtualization on distributed resources for bioinformatics applications. In: eScience. IEEE Fourth International Conference; 2008 Dec; 2008. p. 222–229. Indianapolis, IN, USA.
73. Forer L, Lipic T, Schonherr S, et al. Delivering bioinformatics mapreduce applications in the cloud. In: Information and Communication Technology, Electronics and Microelectronics (MIPRO), 37th International Convention; 2014 May; IEEE; 2014. p. 373–377. Opatija, Croatia.
74. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics.* 2009;25(11):1363–1369.

75. Ahmed L, Edlund A, Laure E, et al. Using iterative MapReduce for parallel virtual screening. In: IEEE 5th International Conference on Cloud Computing Technology and Science 2; 2013 Dec; IEEE; 2013. p. 27–32. Bristol, UK.
76. Zhao J, Zhang R, Zhao Z, et al. Hadoop MapReduce framework to implement molecular docking of large-scale virtual screening. In: Services Computing Conference (APSCC); 2012 Dec; IEEE Asia-Pacific; 2012 Dec. p. 350–353. Guilin, China.
77. Constantine RM, Batouche M. Drug discovery for breast cancer based on big data analytics techniques. In: Information & Communication Technology and Accessibility (ICTA), 5th International Conference; 2015 Dec; 2015. p. 1–6. Marrakech, Morocco.
78. Zou Q, Li XB, Jiang WR, et al. Survey of MapReduce frame operation in bioinformatics. *Brief Bioinform.* 2014;15(4):637–647.
79. Salomoni D, Italiano A, Ronchieri E. WNoDeS, a tool for integrated grid and cloud access and computing farm virtualization. *J Phys.* 2011;331(5):052017.
80. Ronchieri E, Cesini D, D'Agostino D, et al. The WNoDeS cloud virtualization framework: A macromolecular surface analysis application case study. In: Parallel, Distributed and Network-Based Processing (PDP), 22nd Euromicro International Conference; 2014 Feb; IEEE; 2014. p. 218–222. Turin, Italy.
81. Weber R, Gothandaraman A, Hinde RJ, et al. Comparing hardware accelerators in scientific applications: A case study. *IEEE Trans Parallel Distrib Syst.* 2011;22(1):58–68.
82. Couturier R. Designing scientific applications on GPUs. Belfort (France): CRC/Taylor & Francis; 2014.
83. TOP500 list. Available from: <https://www.top500.org>
84. Amdahl GM. Validity of the single processor approach to achieving large scale computing capabilities. In: Proceedings of spring joint computer conference; 1967 Apr; ACM; 1967. p. 483–485. Atlantic City, NJ, USA.
85. Gustafson JL. Reevaluating Amdahl's law. *Commun ACM.* 1988;31(5):532–533.
86. Kasam V, Salzemann J, Botha M, et al. WISDOM-II: screening against multiple targets implicated in malaria using computational grid infrastructures. *Malar J.* 2009 May;1(8):88.
- **This article describes the WISDOM drug discovery challenge.**
87. Breton V, Jacq N, Kasam V, et al. Grid-added value to address malaria. *IEEE Trans Inf Technol Biomed.* 2008;12(2):173–181.
88. Pharmaceutical composition for preventing and treating Malaria, containing compounds that inhibit Plasmepsin II activity, and method of treating Malaria using the same. Available from: <https://patents.google.com/patent/WO2009131384A2/en>
89. Pettifer S, Ison J, Kalas M, et al. The EMBRACE web service collection. *Nucleic Acids Res.* 2010;38(Web Server issue):W683–8.
90. Milanese L. BIOINFOGRID: bioinformatics simulation and modelling based on grid. In: 6th International Workshop on Data Analysis in Astronomy - Modelling and Simulation in Science. 2007 Apr; 2007; p. 178–186. Erice, Italy.
91. Chiappori F, Mattiazzi L, Milanese L, et al. A novel molecular dynamics approach to evaluate the effect of phosphorylation on multimeric protein interface: the alphaB-crystallin case study. *BMC Bioinformatics.* 2016;17(Suppl 4):57.
92. Dove MT, Sullivan LA, Walker AM, et al. Molecular dynamics in a grid computing environment: experiences using DL\_POLY\_3 within the e minerals science project. *Mol Simulat.* 2006;32(12–13):945–952.
93. Harvey MJ, De Britis G. AceCloud: molecular dynamics simulations in the cloud. *J Chem Inf Model.* 2015;55(5):909–914.
94. Addison E, Keinan S. Using quantum molecular design & cloud computing to improve the accuracy & success probability of drug discovery. *Drug Dev Delivery.* 2016;16(2):62–66.
95. Jiang W, Phillips JC, Huang L, et al. Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD. *Comput Phys Commun.* 2014;185(3):908–916.
96. Merelli I, Morra G, D'Agostino D, et al. High performance workflow implementation for protein surface characterization using grid technology. *BMC Bioinformatics.* 2005;6(Suppl 4):S19.
97. Hull D, Wolstencroft K, Stevens R, et al. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.* 2006;34(Web Server issue):W729–32.
98. Oinn T, Li P, Kell DB, et al. Taverna/my Grid: aligning a workflow system with the life sciences community. In: Taylor IJ, Deelman E, Gannon DB, Shields M, editors. *Workflows for e-science.* London: Springer; 2007. p. 300–319.
99. Stevens R, McEntire R, Goble C, et al. myGrid and the drug discovery process. *Drug Discov Today.* 2004;2(4):140–148.
100. Buyya R, Branson K, Giddy J, et al. The virtual laboratory: a toolset to enable distributed molecular modelling for drug design on the world-wide grid. *Concurr Comput Pract Exp.* 2003;15(1):1–25.
101. Zhang W, Du X, Ma F, et al. DGrid: harness the full power of supercomputing systems. In: 5th International Conference Grid and Cooperative Computing Workshops, GCCW'06; 2006 Oct; IEEE; 2006. p. 1–4. Hunan, China.
102. Wang Q, Ye Y, Yu K, et al. A graphical workflow modeler for docking process in drug discovery. In: Chapter 6.7, Grid and Cloud Computing: Concepts, Methodologies, Tools and Applications; Hershey, PA, USA: IGI Global; 2012. p. 1408–1422.
103. Koday MT, Nelson J, Chevalier A, et al. A computationally designed hemagglutinin stem-binding protein provides in vivo protection from influenza independent of a host immune response. *PLoS Pathog.* 2016;12(2):1–23.
104. Fleishman SJ, Whitehead TA, Ekiert DC, et al. Computational design of protein targeting the conserved stem region of influenza hemagglutinin. *Science.* 2011;332(6031):816–821.
- **The paper describes a computational method for designing proteins. The method is applied to the design of different proteins.**
105. Gordon SR, Stanley EJ, Wolf S, et al. Computational design of an  $\alpha$ -gliadin peptidase. *J Am Chem Soc.* 2012;134(50):20513–20520.
106. Wolf C, Siegel JB, Tinberg C, et al. Engineering of kuma030: a gliadin peptidase that rapidly degrades immunogenic gliadin peptides in gastric conditions. *J Am Chem Soc.* 2015;137(40):13106–13113.
107. Das R, Qian B, Raman S, et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins.* 2007;69(58):118–128.
108. Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.* 2008;36(Web Server issue):W233–8.
109. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the rosetta server. *Nucleic Acids Res.* 2004;32(Suppl 2):W526–W531.
110. Lensink MF, Mendez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins.* 2007;69(4):704–718.
111. Bhardwaj G, Mulligan VK, Bahl CD, et al. Accurate de novo design of hyperstable constrained peptides. *Nature.* 2016;538(7625):329–335.
112. Hosseinzadeh P, Bhardwaj G, Mulligan VK, et al. Comprehensive computational design of ordered peptide macrocycles. *Science.* 2017;358(6369):1461–1466.
113. Jayachandran G, Vishal V, Pande VS. Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. *J Chem Phys.* 2006;124(16):164902.
114. Most powerful distributed computing network [Internet]. [cited 2007 Sep 16]. Available from: <http://guinnessworldrecords.com>
115. Beberg A, Ensign D, Jayachandran G, et al. Folding@home: lessons from eight years of volunteer distributed computing. In: 23rd IEEE Parallel & Distributed Processing Symposium; 2009 May; IPDPS; 2009. p. 1–8. Rome, Italy.
- **Authors describe the larger distributed platform worldwide for protein structural analysis.**
116. Kondo D, Javadi B, Malecot P, et al. Cost-benefit analysis of cloud computing versus desktop grids. In: 23rd IEEE Parallel & Distributed Processing Symposium; 2009 May; IPDPS; 2009. p. 1–12. Rome, Italy.
117. Bertis V, Bolze R, Desprez F, et al. Large scale execution of a bioinformatic application on a volunteer grid. In: 22nd IEEE Parallel and Distributed Processing Symposium; 2008 Apr; IPDPS; 2008. p. 1–8. Miami, FL, USA.



22  A. J. BANEGAS-LUNA ET AL.

118. Merelli I, Pérez-Sánchez H, Gesing S, et al. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. *Biomed Res Int*. 2014;2014:134023.
119. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–1303.
120. Zhou W, Li R, Yuan S, et al. MetaSpark: a Spark-based distributed processing tool to recruit metagenomic reads to reference genomes. *Bioinformatics*. 2017;33(7):1090–1092.
121. Niu B, Zhu Z, Fu L, et al. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics*. 2011;27(12):1704–1705.
122. Harnie D, Saey M, Vapirev AE, et al. Scaling machine learning for target prediction in drug discovery using apache spark. *Future Gener Comput Syst*. 2017;67:409–417.
123. Buch I, Harvey MJ, Giorgino T, et al. High-throughput all-atom molecular dynamics simulations using distributed computing. *J Chem Inf Model*. 2010;50(3):397–403.
124. Harvey MJ, Giupponi G, Fabritiis GD. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput*. 2009;5(6):1632–1639.
125. Hess B, Kutzner C, Van der Spoel D, et al. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput*. 2008;4(3):435–447.
126. Marin I, Goga N, Goga M Benchmarking MD systems simulations on the graphics processing unit and multi-core systems. In: IEEE International Symposium on Systems Engineering (ISSE); 2016 Oct; ISSE; 2016. Edinburgh, UK.
127. Hung CL, Chen CC. Computational approaches for drug discovery. *Drug Dev Res*. 2014;75(6):412–418.
128. Vogt M, Bajorath J. Chemoinformatics: a view of the field and current trends in method development. *Bioorg Med Chem*. 2012;20(18):5317–5323.
129. Ma C, Wang L, Xie XQ. GPU accelerated chemical similarity calculation for compound library comparison. *J Chem Inf Model*. 2011;51(7):1521–1527.
130. Malhat MG, El-Sisi AB. Parallel ward clustering for chemical compounds using openCL. In: Tenth International Conference on Computer Engineering & Systems; 2015 Dec; ICCES; 2016. Cairo, Egypt.
131. Lo YT, Wang HW, Pai TW, et al. Protein–ligand binding region prediction (PLB-SAVE) based on geometric features and CUDA acceleration. *BMC Bioinformatics*. 2013;14(Suppl 4):S4.
132. Imbernón B, Cecilia JM, Pérez-Sánchez H, et al. METADOCK: A parallel metaheuristic schema for virtual screening methods. *Int J High Perform Comput Appl*. 2017;32(6):789–803.
133. Yuriev E, Holien J, Ramsland PA. Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. *J Mol Recognit*. 2015;28(10):581–604.
134. Lionta E, Spyrou G, Vassilatis DK, et al. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr Top Med Chem*. 2014;14(16):1923–1938.
135. McIntosh-Smith S, Price J, Sessions RB, et al. High performance in silico virtual drug screening on many-core processors. *Int J High Perform Comput Appl*. 2015;29(2):119–134.
136. Fang Y, Ding Y, Feinstein WP, et al. GeauxDock: accelerating structure-based virtual screening with heterogeneous computing. *PLoS One*. 2016;11(7):e0158898.
137. Feinstein W, Brylinski M. Structure-based drug discovery accelerated by many-core devices. *Curr Drug Targets*. 2016;17(14):1595–1609.
138. Krige S, Mackey M, McIntosh-Smith S, et al. Porting a commercial application to openCL. In: Proceedings of the International Workshop on OpenCL 2013 & 2014; 2014 May; IWOCCL '14; 2014. Bristol, UK.
139. Harvey MJ, Giupponi G, Fabritiis GD. ACEMD. Accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput*. 2009;5(6):1632–1639.
140. Sukhwani B, Herboldt MC. Fast binding site mapping using GPUs and CUDA. 2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW); 2010 Apr; IPDPSW; 2010. Atlanta, GA, USA.
141. McArt DG, Bankhead P, Dunne PD, et al. cudaMap: a GPU accelerated program for gene expression connectivity mapping. *BMC Bioinformatics*. 2013;14:305.
142. Hu X, Liu Q, Zhang Z, et al. SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. *Cell Res*. 2010;20(7):854–857.
143. Irwin JJ, Shoichet BK, Mysinger MM, et al. Automated docking screens: a feasibility study. *J Med Chem*. 2009;52(18):5712–5720.
144. Tsai TY, Chang KW, Yu-Chian Chen C. iScreen: world's first cloud-computing web server for virtual screening and de novo drug design based on TCM database@Taiwan. *J Comput Aided Mol Des*. 2011;25(6):525–531.
145. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res*. 2011;39(Web Server issue):W270–7.
146. Prakhov ND, Chernorudskiy AL, Gainullin MR. VSDocker: a tool for parallel high-throughput virtual screening using AutoDock on windows-based computer clusters. *Bioinformatics*. 2010;26(10):1374–1375.
147. Sánchez-Linares I, Pérez-Sánchez H, Cecilia JM, et al. High-throughput parallel blind virtual screening using BINDSURF. *BMC Bioinformatics*. 2012;13(Suppl 14):S13.
148. Zoete V, Daina A, Bovigny C, et al. SwissSimilarity: a web tool for low to ultra high throughput ligand-based virtual screening. *J Chem Inf Model*. 2016;56(8):1399–1404.
149. Shang J, Dai X, Li Y, et al. HybridSim-VS: a web server for large-scale ligand-based virtual screening using hybrid similarity recognition techniques. *Bioinformatics*. 2017;33(21):3480–3481.
150. Li H, Leung KS, Wong MH, et al. USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. *Nucleic Acids Res*. 2016;44(W1):W436–W441.
151. Wang X, Chen H, Yang F, et al. iDrug: a web-accessible and interactive drug discovery and design platform. *J Cheminform*. 2014;6:28.
152. Gong J, Cai C, Liu X, et al. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics*. 2013;29(14):1827–1829.
153. Koes DR, Camacho CJ. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res*. 2012;40(W1):W409–W414.
154. Dong J, Yao ZJ, Wen M, et al. BioTriangle: a web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions. *J Cheminform*. 2016;8:34.
155. Click2Drug. Swiss Institute of Bioinformatics. Available from: <http://www.click2drug.org>
156. Novartis Case Study. Amazon web services [Internet]. Available from: <https://aws.amazon.com/it/solutions/case-studies/novartis/>
157. List M. Using docker compose for the simple deployment of an integrated drug target screening platform. *J Integr Bioinform*. 2017;14:2.
158. Mark S, Meier MS, Toth DM. Accelerating AutoDock Vina with Containerization. In Proceedings of the Practice and Experience on Advanced Research Computing, p. 36. ACM, 2018. Available from: <https://dl.acm.org/citation.cfm?id=3219154Hisle>
159. AstraZeneca taps AI for drug discovery in deal with Berg. Reuters. 2018. Available from: <https://www.reuters.com/article-us-astrazeneca-ai-berg/astrazeneca-taps-ai-for-drug-discovery-in-deal-with-berg-idUSKCN1B81G1>

### 3.3 A REVIEW OF LIGAND-BASED VIRTUAL SCREENING WEB TOOLS AND SCREENING ALGORITHMS IN LARGE MOLECULAR DATABASES IN THE AGE OF BIG DATA

<b>Título</b>	A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data
<b>Autores</b>	Banegas-Luna AJ, Cerón-Carrasco JP, Pérez-Sánchez H.
<b>Revista</b>	Future Medicinal Chemistry
<b>Año</b>	2018
<b>Volumen</b>	10 (22)
<b>Páginas</b>	2641-2658
<b>Estado</b>	Publicado
<b>DOI</b>	<a href="https://doi.org/10.4155/fmc-2018-0076">https://doi.org/10.4155/fmc-2018-0076</a>
<b>IF(2017)</b>	3.969
<b>Categoría</b>	Chemistry, Medicinal, 9/59, Q1

#### Contribución del Doctorando

El doctorando Antonio Jesús Banegas Luna declara ser el autor principal y contribuyente principal del artículo *A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data*.

Review

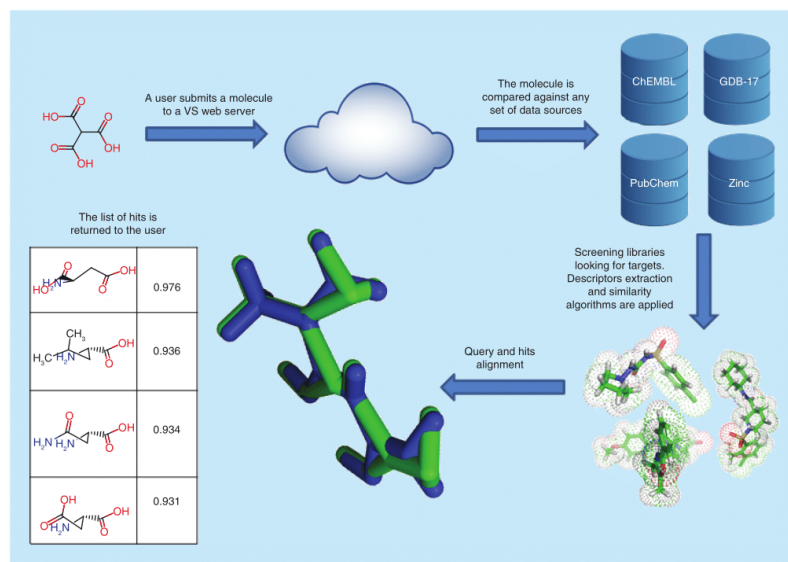
For reprint orders, please contact: [reprints@future-science.com](mailto:reprints@future-science.com)Future  
Medicinal  
Chemistry

## A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data

Antonio-Jesús Banegas-Luna<sup>\*1</sup>, José P Cerón-Carrasco<sup>1</sup> & Horacio Pérez-Sánchez<sup>1</sup><sup>1</sup>Bioinformatics & High Performance Computing Research Group (BIO-HPC), Computer Engineering Department, Universidad Católica San Antonio de Murcia (UCAM), Campus de los Jerónimos, 30107, Murcia, Spain<sup>\*</sup>Author for correspondence: [ajbanegas@alu.ucam.edu](mailto:ajbanegas@alu.ucam.edu)

Virtual screening has become a widely used technique for helping in drug discovery processes. The key to this success is its ability to aid in the identification of novel bioactive compounds by screening large molecular databases. Several web servers have emerged in the last few years supplying platforms to guide users in screening publicly accessible chemical databases in a reasonable time. In this review, we discuss a representative set of online virtual screening servers and their underlying similarity algorithms. Other related topics, such as molecular representation or freely accessible databases are also treated. The most relevant contributions to this review arise from critical discussions concerning the pros and cons of servers and algorithms, and the challenges that future works must solve in a virtual screening framework.

### Graphical abstract:



First draft submitted: 12 March 2018; Accepted for publication: 28 September 2018; Published online: 30 November 2018

**Keywords** chemical database • chemical space • descriptors • molecular fingerprints • molecular representation • pharmacophore modeling • similarity searching • virtual screening • web servers

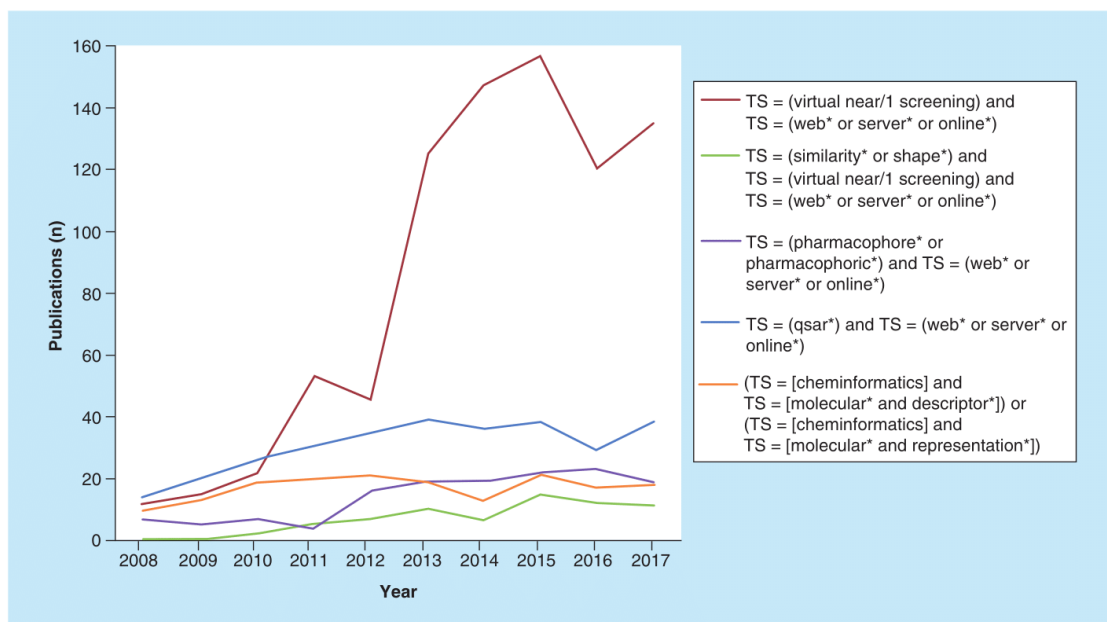
newlands  
press

Designing a new drug is a long and expensive process, usually costing billions of dollars and may well take >10 years [1]. While a few decades ago most of the investment went on experimental research to identify novel drug candidates in the hope that one would show biological activity against a target protein [2], nowadays computer-aided drug discovery (CADD) tools have gained importance because they can reduce the number of ligands that need to be screened in experimental assays. Among existing CADD techniques, virtual screening (VS) is one of the most important and widely used [3]. VS scans large databases of compounds, either of known molecules or of molecules that could be synthesized, searching for those that have the highest probability of showing some property of interest, for example, a beneficial effect on cholesterol levels [4]. Although smaller databases can also be screened, the present contribution mainly focuses on tools for screening large databases, which is one of the predominant cases. The success of VS is defined in terms of finding lead-like hits with potential biological activity against the intended drug targets rather than several hits, which may make it a reliable, profitable and time-saving technique [5].

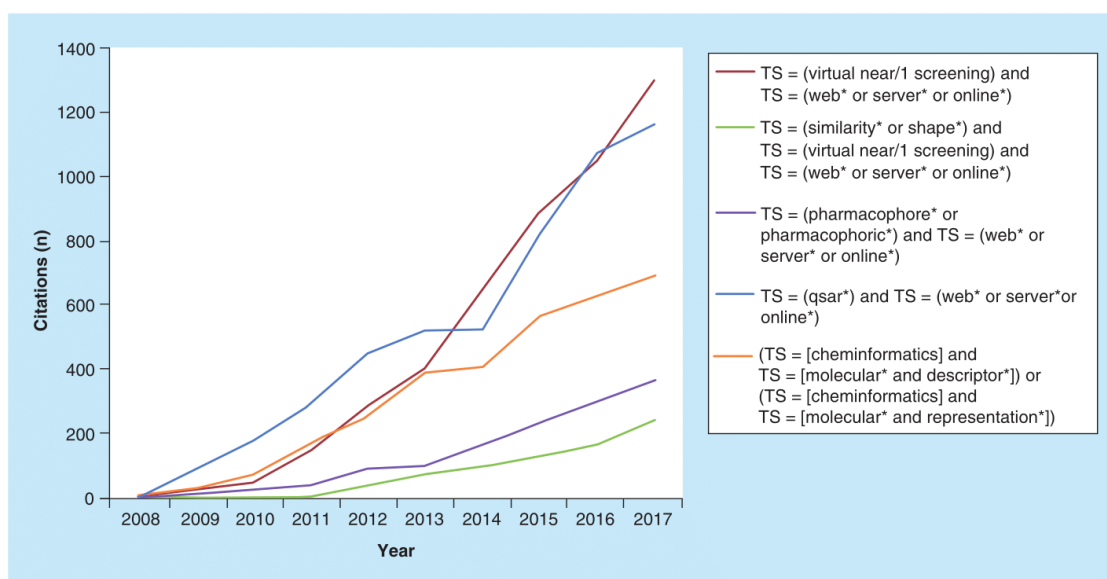
VS techniques are usually classified into two major categories: structure-based (SBVS) and ligand-based (LBVS). VS SBVS, which is more suitable for finding structurally novel ligands, is the preferred method when the 3D structure of the target protein has been characterized experimentally (e.g., x-ray crystallography [6], NMR [7,8] or homology modeling [9,10]). In cases where the 3D structure of the target is unknown or its prediction by structure-based methods is challenging, LBVS is the preferred protocol [11–13]. The functionality of the core engine of LBVS is based on the assumption that molecules with a similar structure (in terms of structure, pharmacophoric features, molecular fields, etc.) also exhibit a similar behavior. Thus, LBVS relies on a comparison of the structure of molecules. The choice of one or other of the approaches strongly depends on the case in question. Generally, LBVS techniques, such as substructure mining and fingerprint searches, are faster than those of SBVS (e.g., molecular docking). Indeed, they have proved to be reliable in many cases for finding promising compounds based on physical, chemical or thermodynamic properties [14], which is why this contribution is mainly focused on LBVS methods.

Ligand-based methods depict compounds through a set of descriptors that represent molecular characteristics in numerical form. Descriptor selection is important for many reasons, but specifically, because the correct selection of descriptors can lead to cost-effective models and reduce the noise of redundant descriptors. Descriptors can be classified according to several criteria, such as the nature of the characteristics they represent – for example, they could be categorized as geometrical, topological, thermodynamic, constitutional or electronic [15] – among others. Another frequent classification is based on the geometrical structure of a molecule, being grouped as 1D, 2D or 3D [16]. On the basis of these and other criteria, many software tools, such as Omega [17], OSIRIS [18] or MoKa [19,20], can help in computing molecular descriptors and provide a well classified set of the same. Unfortunately, although such tools can calculate the same descriptors, their estimated values may not be the same leading to dissimilar evaluations of similarity. Consequently, the selection of one or the other might have an impact on the final output of the screening. In addition to difficulties involving the complexity of descriptor calculations and the selection of the most suitable representation, LBVS must handle the overwhelming volume of information available in several chemical databases. Large amounts of molecular and biological activity data are available nowadays through freely accessible databases, such as ZINC [21], PubChem [22] or ChEMBL [23]. The sizes of such databases (Table 1), ranging from a few thousands in Drug Bank [24] to almost one hundred million in PubChem, point to the quantity and diversity of the data sources available for screening.

Chemical databases fit the requirements to be considered as Big Data resources including the 5Vs (volume, variety, velocity, veracity and value) rule, which is accepted as one of the most extended standards to define big resources [32]. In a CADD context, the requirements of volume, variety, velocity, veracity and value can be seen in molecules, algorithms and techniques [33–35]. As the screenable databases are very large and provide such a diversity of information, sometimes of limited or unknown usefulness, heuristics (e.g., a preliminary filtering of compounds) are typically used to avoid spending excessive time on unpromising calculations. To handle all these issues, technology has been extensively applied to automate some steps of LBVS (e.g. descriptor calculation, database screening and molecular comparison) [36–38]. Because of this, the importance of LBVS web servers has grown in the last decade (Figures 1 & 2) resulting in an increasing number of publications regarding related topics [39–42]. However, the lack of a critical assessment of the existing servers and the impact of molecular representations on their performance is an important issue that needs to be covered. To help fill this gap, this work identifies the main implementation details of the existing web servers and provides a critical overview of this topic.



**Figure 1.** Number of publications indexed by the Web of Knowledge concerning ligand-based virtual screening web servers. The asterisks (\*) are a wildcard symbol which means 'any character(s)' in the context of a Web of Science query. TS: Topic searched; QSAR: Quantitative structure–activity relationship.



**Figure 2.** Number of citations indexed by the Web of Knowledge concerning ligand-based virtual screening web servers. The asterisks (\*) are a wildcard symbol which means 'any character(s)' in the context of a Web of Science query. TS: Topic searched; QSAR: Quantitative structure–activity relationship.

Table 1. Some of the most well-known, freely-accessible databases used in virtual screening (accessed on 16 September 2018).

Database	Compounds	Website	Ref.
GDB-17	166.4 billion	<a href="http://gdb.unibe.ch">http://gdb.unibe.ch</a>	[25]
PubChem	95.4 million	<a href="https://pubchem.ncbi.nlm.nih.gov">https://pubchem.ncbi.nlm.nih.gov</a>	[22]
ChemSpider	65 million	<a href="http://chemspider.com">http://chemspider.com</a>	[26]
ZINC	35 million	<a href="http://zinc.docking.org">http://zinc.docking.org</a>	[21]
ChemMine	6.2 million	<a href="http://chemminedb.ucr.edu">http://chemminedb.ucr.edu</a>	[27]
ChEMBL	1.7 million	<a href="https://ebi.ac.uk/chembl">https://ebi.ac.uk/chembl</a>	[23]
ChemBank	1.7 million	<a href="http://chembank.broadinstitute.org">http://chembank.broadinstitute.org</a>	[28]
Binding DB	648,871	<a href="https://bindingdb.org">https://bindingdb.org</a>	[29]
Protein Data Bank	140,109	<a href="http://rcsb.org">http://rcsb.org</a>	[30]
Drug Bank	10,562	<a href="http://drugbank.ca">http://drugbank.ca</a>	[24]
KEGG Ligand	5645	<a href="http://genome.jp/kegg/drug">http://genome.jp/kegg/drug</a>	[31]

The first part of this review introduces the most relevant topics for a better understanding of the servers described afterwards, for example, 2D and 3D comparison algorithms, filters and descriptor extraction. The second half presents the most widely used online servers that help users to perform LBVS. Finally, an overview of the main advantages and drawbacks of every server and technique is given in Section 4, identifying the challenges that remain in relation with this emerging subject.

### Molecular representations in LBVS

The representation of molecules using models is the initial input of the first LBVS step. Such models represent a collection of molecular properties, which are classically classified into 1D, 2D or 3D subgroups [43]. Each representation has a set of characteristics that make it especially helpful in some steps of LBVS. Calculating similarity is an example of a process that is strongly impacted by the dimensionality of the representation chosen. Table 2 summarizes the core equations for calculating the similarity between a query and a candidate ligand, together with the required dimensionality and a few examples of where they are applied.

The following sections introduce different types of molecular representations and emphasize their impact in LBVS. A set of representative comparison algorithms and screening filters are also presented.

### Representing a molecule with descriptors

Descriptors or molecular properties are mainly used to build a fingerprint representing a molecule. Such fingerprint may contain 2D and 3D descriptors, which can be represented as binary or real values. The election of 2D or 3D descriptors is a crucial choice in the fingerprint building process, especially considering that not all the descriptors have a relationship with the biological activity. Molecules of similar 3D shape and with similar properties could share biological activities, even if their 1D and 2D representations are not similar, because the binding affinity between molecules and target proteins is dependent on atomic interactions in the 3D space [16]. This may lead to 3D descriptors being considered a better choice; however, a single molecule usually has many 3D representations, which makes 3D models more demanding in terms of storage space and calculation time [15,53]. By contrast, although 2D descriptors perform well, they omit the characteristics related with the spatial disposition of the molecule (e.g., shape or atom distribution), whereas 3D descriptors correctly account for such properties when evaluating similarity [16]. Many software packages provide descriptor calculation functionalities (Table 3), but not all of them extract the same set of descriptors, so that the correct choice of software will strongly depend on the particular needs.

### Implementation of 2D representations

As already mentioned, 2D representations are a cost-effective solution in many cases. Fingerprints are a very common way of representing molecules in LBVS, but there is a large variety of 2D molecular models including graphs and vectors. The 2D representation of molecules in graph form is simple and helps in the construction of graph-based molecular models, which are especially useful for representing the connections between atoms. Such models are able to include most of the information represented by descriptors for QSAR (quantitative structure–activity

Table 2. Summary of molecular similarity theoretical background.			
Approach	Dimension	Similarity formula	Ref.
2D fingerprints 2D/3D fingerprints combination	2D	$D(Q, C) = \frac{ Q \wedge C }{ Q \vee C }$	[44,45]
Physical-chemical properties Reverse SBVS/LBVS	3D	$D(Q, C) = \sum_{i=1}^n \sqrt{Q_i^2 - C_i^2}$	[46,47]
Align-IT 3DAPfp/3DXfp Topological analysis	3D	$D(Q, C) = \sum_{i=1}^n  Q_i - C_i $	[48,49]
PharmShapeCC	3D	$D(Q, C) = \frac{\sum_i (Q_i \wedge C_i)}{\sum_i (Q_i \vee C_i)}$	[50]
Avalanche	3D	$D(Q, C) = \sum_{i=1}^{16} w_i \sum_{j=1}^{Nbins}  n_{ij}^Q - n_{ij}^C $	[51]
Machine learning	3D	$D(Q, C) = \frac{\sum_{i=1}^m \min(x_{Q,i}, x_{C,i})}{\sum_{i=1}^m x_{Q,i} + \sum_{i=1}^m x_{C,i} - \sum_{i=1}^m \min(x_{Q,i}, x_{C,i})}$	[52]

C: Candidate ligand; D: Similarity; LBVS: Ligand-based virtual screening; Q: Query molecule; SBVS: Structure-based virtual screening.

Table 3. Examples of software packages for calculating descriptors.			
Software	Descriptors	Status	Ref.
DRAGON	4885	Commercial	[54]
MODEL	3778	Freeware	[55]
ChemoPy	3735	Freeware	[56]
PADEL	1875	Freeware	[57]
CODESSA	1500	Commercial	[58]
ADRIANA. Code	1244	Commercial	[59]
MOE	300	Commercial	[60]
CDK Descriptor Calculator	50	Freeware	[61]
RDKit	48 <sup>†</sup>	Freeware	[62]

<sup>†</sup> A total of 40 descriptors and 8 types of fingerprints.

relationship) without much complexity, so that graphs are suitable for the rapid development of graph-matching algorithms to evaluate similarity [63]. A more generic approach, however, is to use vectors to represent molecular properties, such as MW or the number of rotatable bonds. Each position in the vector represents a descriptor. Vectors provide two main benefits: they are suitable for representing 2D and 3D properties; and calculating similarity ( $S$ ) between two binary vectors ( $Q$  and  $C$ ) is extremely easy and fast with Boolean logic (Equation 1) [16].

$$S(Q, C) = \frac{|Q \wedge C|}{|Q \vee C|} \quad (\text{Equation 1})$$

A fingerprint is a particular case of binary vector, where each position contains 1 or 0 to represent the presence or the lack of a descriptor. The size of this sort of fingerprint grows with the number of descriptors represented but, in order to use the space efficiently, they can be compressed into shorter ones [44]. Due to their small size, fingerprints can easily be calculated in advance and stored in a relational database to be retrieved later.

All the introduced methods are useful for modeling molecules, but fingerprints are the most frequently used because of their computational efficiency and proven effectiveness [64,65].

#### Implementation of 3D fingerprints

As mentioned above, fingerprints can represent any type of 3D descriptor, and several implementations of 3D fingerprints can be found in the literature. For example, PMIfp [66] is a 3-bit scalar fingerprint that measures the principal moments of inertia scaled to MW. An alternative approach is USR [67], which in a 12-bit print that represents the Euclidean distance distributions calculated with respect to four reference points. An extension of USR is USRCAT [68], which divides the atoms into five categories to extend the original print to a 60-bit one. To study the impact of stereochemistry on LBVS, the novel 3DAPfp and 3DXfp fingerprints (of 16 and 80 bits, respectively) were computed with the JChem library. Both fingerprints represent an extension of their 2D counterparts amplified with stereochemical properties [48]. It should be mentioned here that not only properties directly extracted from the molecule are allowed in a fingerprint but also customized descriptors, an aspect that was used to develop an algorithm based on molecular topological analysis [49]. Contrary to the other approaches, molecular topological analysis merges descriptors and energy charges. Among the 3D fingerprints presented in this section, 3DAPfp and 3DXfp showed the best performance for screening large datasets. Whatever the case, a customized collection of descriptors can be used to build a fingerprint that suits specific needs.

#### Pipelining 2D & 3D representations

Although 2D- and 3D-based representations have been presented separately, they may be applied in the same workflow to complement each other. This approach has already been used to develop a pipeline for evaluating potential PDE4 and PDE5 inhibitors [69]. A 2D shape similarity is conducted at the early stage, which is complemented with a filter by range and a diversity study. Finally, the best-scored compounds are compared in a more detailed 3D LBVS step.

Just as 2D and 3D representations are not mutually exclusive, SBVS and LBVS can also work together in a reverse pipeline [47]. Contrary to the usual trend, SBVS is applied first to find the three best scoring small fragments, whose fingerprints will be used later as queries. Next, LBVS screens the database searching for those molecules whose fingerprints are similar enough to the three selected, according to Euclidean distance. Since this method deals with fragments rather than with full ligands, it allows the user to focus on specific regions within the chemical space. Nevertheless, the SBVS stage may perform worse than other 2D fingerprint approaches (e.g., turbo similarity [70], FTrees [71]).

#### Algorithms for molecular comparison

While the screening of many different data sources is the most usual scenario, not all comparison algorithms perform well in this situation. However, PharmShapeCC is able to screen trillions of compounds from thousands of combinatorial libraries in parallel. The key point of such a good performance is that the libraries are synthesized based on the assumption that the binding poses of the active members of each library do not vary importantly. Next, the PharmShape [50] tool, which carries out 3D pharmacophore and shape screening, is applied to individual libraries to find the best candidates.

Avalanche is another tool that can carry out extremely fast shape/feature-based comparisons to determine four molecular surfaces, taking the Connolly surface [72] as the reference one. The Connolly surface, which is especially relevant because of its high degree of accuracy, is defined as the surface obtained by rolling up a sphere of radius 'r' on the van der Waals surface. In this approach, fingerprints are replaced by histograms representing many physical and chemical properties (e.g., molecular volume, distribution of hydrogen bond donor/acceptors on the molecular surface), which are used to compare both molecules.

In the recent years several machine-learning approaches, such as Bayesian networks, neural networks, random forest or genetic algorithms have been introduced in a variety of contexts related with medicine and bioinformatics [73–75]. The key principle of these approaches is that the system learns from data through a learning process, which is usually classified as supervised or unsupervised. Some of such techniques have been combined in a pipeline [52],



which replaces fingerprints by a so-called signature, to describe the way that atoms are connected in a molecule. In this approach, genetic algorithms and support vector machines create a training model that is applied to screening the database in question. A principal component analysis is also conducted during the model creation process to create a pool of features, which will help to optimize the model.

#### Excluding less promising compounds through filters

Even though fingerprints can compare two molecules very quickly, LBVS would still be very slow if the query were compared against every molecule in the database, due to the overwhelming size of the current databases. Consequently, filters are frequently used, using descriptors aiming at reducing the number of compounds in subsequent LBVS steps while discarding bad candidates.

Lipinski's rule of five, which has recently been revisited using pharmacokinetic data in rat [76], is a typical filter and extensively reported in the literature [77–79]. The prediction of pharmacokinetic-related properties can provide valuable information for decision making, because many of the compounds predicted as drugs fail in late stages due to toxicity-related issues. Absorption, distribution, metabolism & toxicity are the properties typically evaluated to identify toxicity problems [80]. On the other hand, some compounds are likely to interfere in experimental screening techniques, mainly as a result of their potential reactivity leading to false positives. These undesired molecules are usually referred to as pan-assay interfering substances (PAINS) and should be excluded from bioassays [81]. There are many other filters that can be applied to predict compounds less prone to fail in further stages, such as Veber [82] and 3/75 [83]. All these filters can be easily implemented using a relational database to make them suitable for web applications, and they are frequently combined in the same experiment to avoid wasting time on lead candidates that would be toxic or metabolized by the body into an active form [84–86].

#### Web servers

The collection of web servers providing VS services has grown in recent years [37], and their effectiveness is usually assessed in terms of the accuracy of the results and calculation speed. The application of appropriate filters and the calculation of efficient representations, as introduced in Section 2, are key points for achieving a good balance between accuracy and speed [51,87]. Here, we present a representative set of high-throughput LBVS servers (Table 4) that use different techniques which, either separately or in combination with others, involve very different calculation times, ranging from seconds to days. The main features of each server, either reported in the original paper or observed on the website, are summarized in the following table.

Similarity searching is the most rapid and straightforward LBVS approach to search for compounds that are chemically or physicochemically similar to the query molecule [103,104]. Superimposé, which is a wizard-style server, provides not only similarity searching but also binding-site searching. Searches are classified into three levels depending on the size of the molecules: small molecule level, macromolecule based on substructures level and protein level. Specific options are proposed for each level, including dedicated libraries and comparison algorithms. To perform similarity searching on small molecules, Superimposé supports two 3D comparison algorithms: score1 and sd.best\_compare. The former is a 3D cyclical algorithm that applies a branch-and-bound technique on a pair of new artificial molecules obtained by removing some atoms from the original ones. Atoms are removed while the original molecules and the new ones remain spatially similar. The latter is a two-steps algorithm to find similar molecules with different connection schema. The first step performs a normalization of the atoms independently of any rotations or transformations, whereas the second superimposes and refines the alignment of the atoms cyclically. Other servers implementing 3D shape similarity are available on the web, such as wwLigCSRre and AURAmol. The idea behind both servers remains the same but wwLigCSRre uses LigCSRre to assess similarity, whereas AURAmol implements the maximum common subgraph algorithm. As regard to the databases that can be screened, while AURAmol is limited to its own set of compounds, wwLigCSRre can screen ChemBridge, Drug Bank and many other datasets.

To cover the 2D and 3D screening of ultralarge libraries, SwissSimilarity carries out similarity searching, allowing users to choose among many similarity algorithms and libraries of small molecules. A 2D fingerprint-based search is provided for all the libraries, including the largest ones. There are also four 3D algorithms available – Electroshape-5D [105], Spectrophores [106], Shape-IT [107] and Align-IT [107]. However, to preserve the server responsiveness, such algorithms cannot screen the largest datasets. On the one hand, Electroshape-5D and Spectrophores, which use city-block distance [108] to assess similarity, represent a nonsuperpositional shape-based approach. On the other hand, Shape-IT, which is a superpositional shape-based method, and Align-IT, which is pharmacophore based,

Review Banegas-Luna, Cerón-Carrasco &amp; Pérez-Sánchez

Website	Pros	Cons	MC	Technologies			Ref.
				Screenable libraries	Software	Similarity	
AURAmol http://bit.ly/2f1tHiX	Multiplatform Molecule format storable in database	Must know cutoffs <i>a priori</i> Existence of more effective algorithms. Scales bad with the size of database		In-house DB	C/C++	MCS	[88–90]
Superimposé†	Diversity of libraries and algorithms Target proposal	Low performance	2540	SuperDrug, NCI, Ligand Depot	JMol, STRAP	Score1, sd_best.compare	[91]
wwLigCSRre http://bit.ly/2fuFkja	Custom database Focused libraries Regular expressions	Limited library size Default regular expressions are improvable	50	Drug Bank, Chembridge, CPDB, Focused libs	JMol	LigCSRre	[92]
ZINCPharmer http://bit.ly/2gdhL0b	Precalculated conformers Usage of indexes	Only screens ZINC Runtime scales with the number of filters	10	ZINC	JMol	Pharmer	[93]
ChemMapper†	Modular design	Static cutoffs No consensus	50	ChEMBL, Drug Bank, Binding DB, KEGG, PDB	JMol	SHAFTS, USR, 2D fingerprint	[94]
iDrug http://bit.ly/2eCCO19	Custom database Optional flexibility	Static cutoffs Low performance		NCI, ZINC	Java, MySQL, Python, JSP	SHAFTS	[95]
LBVS†	Custom database Different validation methods	Static cutoffs Risk of overfitting		BindingDB, ChEMBL	ChemDoodle, D3.js, MongoDB, MySQL	Bayesian learning	[96]
UFSRAT http://bit.ly/2f1sF6y	Effective descriptor storage LBVS and docking	Does not visualize hits Small databases		EDULISS		USR	[97,98]
SEABED†	Format supporting Docking/QSAR Custom database	Time consuming Risk of overfitting				Machine learning 1D/2D fingerprints	[99]
USR-VS http://bit.ly/2eCcg0t	Multithreading Preload in memory	Only screens ZINC Consumes memory	35	ZINC	lview	USR, USRCAT	[100]
SwissSimilarity http://bit.ly/2fWdQ4l	Several well-defined libraries Friendly interface Multithreading	Large libraries not screenable with 3D methods	20	Up to 30 databases		2D fingerprint, Electroshape-5D, Spectrophores, Shape-IT, Align-IT	[101]
HybridSim-VS http://bit.ly/2Ddjs8T	Analysis options Many screenable datasets Use of 2D/3D fingerprints	No cutoff available Static datasets Few configuration options		Drug Bank, focused libraries, TCM, commercial libraries		FP2 + WEGA MACCS + WEGA	[102]

† Temporarily unavailable.  
 CPDB: Carcinogenic potency database; EDULISS: Edinburgh University ligand selection system; KEGG: Kyoto encyclopedia of genes and genomes; LBVS: Ligand-based virtual screening; MACCS: Molecular access system; MC: Maximum number of conformers; MCS: Maximum common subgraph; NCI: National Cancer Institute; QSAR: Quantitative structure–activity relationship; TCM: Traditional Chinese medicine; VS: Virtual screening.

score their results using the Tanimoto coefficient. Additionally, a combined 2D/3D screening method, which implements a consensus scoring function combining Tanimoto coefficient and city-block distance, is available. Contrary to SwissSimilarity, HybridSim-VS is a server that evaluates shape-based similarity by combining 2D fingerprints and WEGA, which is a 3D shape similarity software. Although both servers are similar in terms of the

screenable datasets that they offer, HybridSim-VS always calculates 3D similarity to take advantage of geometric information.

As an alternative to similarity searching, pharmacophore searching is another relevant LBVS technique. On the basis of pharmacophore modeling, the ZINCPharmer server screens a library of conformations calculated from a subset of compounds purchasable from ZINC. It uses an indexed database system to store the pharmacophore features, which speeds up the searches, and provides a query editor for users to refine their searches. Its main limitation is that it is closely coupled to ZINC and the library provided needs to be synchronized periodically. ZINCPharmer relies on Pharmer [109] as the core algorithm to create the initial query, which may be calculated from a ligand, a ligand–protein interaction, a protein–protein inhibitor interaction or third party software. The performance of Pharmer is scaled in accordance with the breadth and complexity of the query, not with the size of the library, but it has shown good performance at generating pharmacophore queries [109] and in combination with other methods [110].

To offer both shape and pharmacophore searching, USR-VS screens a large dataset of 3D conformers obtained from ZINC and has demonstrated its excellent performance. It loads all the input conformers in the memory and takes advantage of multi-threading to compare many of them against the query at the same time. As similarity algorithms, USR and USRCAT, which represent the shape and pharmacophore approaches respectively, are supported. Despite the good performance of USR, both USR-VS and ZINCPharmer share the limitation of screening only ZINC compounds. The UFSRAT server takes its name from the UFSRAT algorithm, which is an extension of USR that complements the typical shape similarity with information about the electrostatics of the atoms. Unlike the above-mentioned servers, UFSRAT can screen several databases (e.g., ChemBridge and MayBridge) with up to 4.8 million compounds. Similarly to USR-VS, the iDrug server provides both similarity and pharmacophore searching. It manages 3D representations to take advantage of superimposition and has SHAFTS [111] as the underlying similarity algorithm. SHAFTS is a suitable software for large-scale VS with bioactive compounds, and has been successfully applied in many studies, including that which led to the discovery of p90 ribosomal S6 protein kinase inhibitors and the development of novel B-Raf<sup>G609E</sup>-selective inhibitors. It has also been used to analyze the bioactivity of green tea [112–115]. ChemMapper exhibits some common features with iDrug, including the use of SHAFTS and superimposition. However, other 2D (fingerprints and MACCS [116]) and 3D (USR) similarity methods are also available. The main feature that differentiates ChemMapper from others is that it offers the possibility of screening a custom library. Additionally, from a software point of view, it is clearly split into the following five components:

- A chemical database;
- An in-house similarity searching method;
- A compound–target annotation database;
- A network inference method for target recommendations;
- A viewer tool for visualizing results.

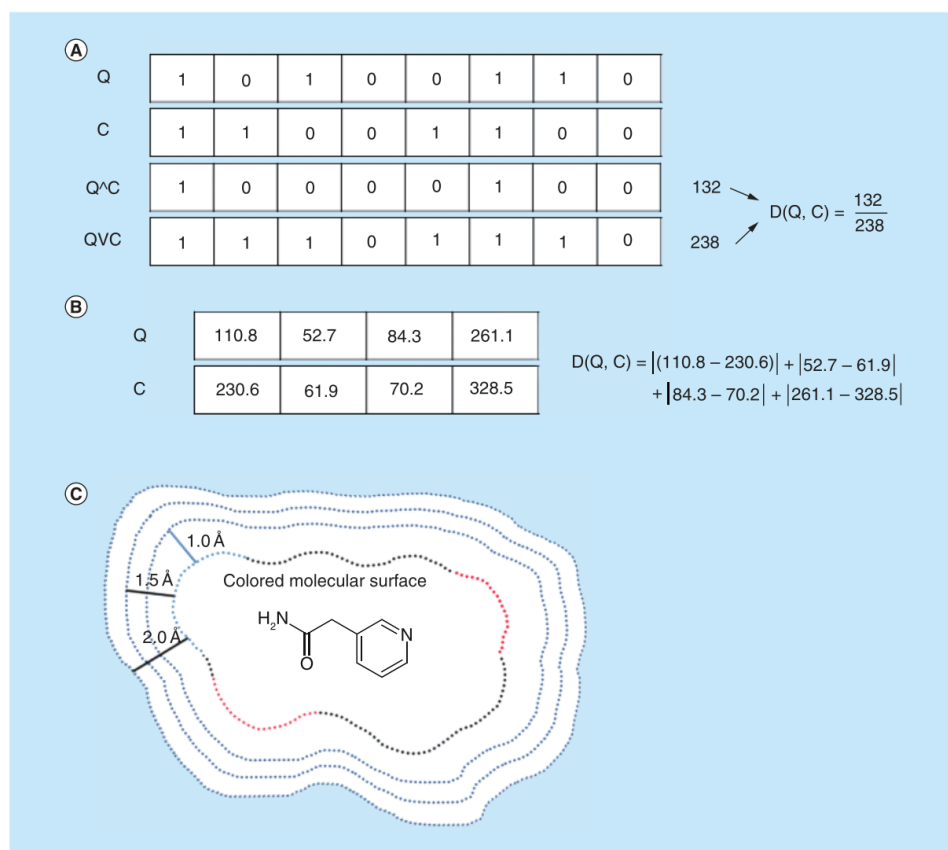
Along with shape and pharmacophore similarity, QSAR modeling is the other major ligand-based approach. SEABED and LBVS are examples of web servers applying this technique. SEABED can run QSAR, docking or combined jobs. In addition, it also exhibits other features such as receptor preparation, library editing or VS on protein mutants. The server supports QSAR studies using a large list of 1D and 2D fingerprints, and implements several machine-learning techniques, including naive Bayes and random forest. SEABED can choose the compounds to screen, but the users may upload their own databases. LBVS is an alternative server using Bayesian learning for building QSAR models. As with SEABED, users can screen their custom datasets or finding a target in BidingDB or ChEMBL. This server is capable of dealing with big data resources and can be used for lead identification and optimization.

## Discussion

### Molecular representations

Section 2 presented a set of molecular representations typically used in LBVS and some molecular comparison algorithms. Here we discuss the positive and negative aspects of such tools. Figure 3 illustrates examples of how fingerprints and Avalanche perform comparisons.

Review Banegas-Luna, Cerón-Carrasco &amp; Pérez-Sánchez



**Figure 3. Examples of molecular similarity assessments using fingerprints and histograms. (A)** Comparison of binary 2D fingerprints. **(B)** Comparison of 3D fingerprints using city-block distance. **(C)** Elections of 16 points on the molecular surface to build histograms in Avalanche  
Reproduced with permission from [51] © Springer International Publishing Switzerland (2015).

The 2D representations are easy and fast to compute similarity. Graph-based similarity reported a high success rate in classification ( $\sim 70\%$ ) [63]. In such study, the results indicated that the accuracy was increased when information about the bonds and their connectivity was included in the graph. This happened because the type of bond is an important factor for determining the characteristics of a molecule. The efficient extraction of appropriate features to calculate similarity remains the main challenge in the construction of molecular graphs. On the other hand, a flat representation with binary fingerprints may help to screen large datasets. In terms of speed of similarity calculations, such fingerprints were seen to perform 20-times faster than the implementation provided by OpenEye [117] and 10-times faster than single-instruction, multiple-LINGO (SIML) [118] when screening a dataset of  $2^{15}$  molecules [44]. However, the main drawback of this representation is that its computational requirements grow proportionally with the size of the database (e.g., space on disk and memory usage). In an attempt to overcome the limitations of 2D methods, 3D representations represent the alternative approach. USR was able to screen 3.5 billion molecules in  $<12$  min on one single processor, which was  $>2000$ -times faster than the fastest method reported at that time [67]. The success of USR lies on how it encodes molecular shapes with descriptors. Although such descriptors only need to be calculated once, they must be rigorously selected to reach a high efficiency. Both USR and USRCAT, performed slightly better than ElectroShape when they were benchmarked against Directory of Useful Decoys, Enhanced (DUD-E) [68]. The performance of both methods is highly dependent on the query molecule and the

library being screened, and so their performance could be improved by adjusting some parameters to the query molecule. The topological analysis approach was also tested against DUD-E [49]. In this case, the enrichment factor was much better than USR (in 83 out of 99 datasets) and USRCAT (in 42 out of 99 datasets). This can be explained by the fact that it is designed to cover the weaknesses of the other methods. A last group of 3D fingerprints, 3DAPfp and 3DXfp, was compared in terms of LBVS performance against other 3D fingerprints (PMIfp, USR and USRCAT) and their corresponding 2D implementations (APfp and Xfp). The tests showed that 3DAPfp was the best method for representing 3D shapes, and it outperformed APfp [48]. Nonetheless, 3DXfp was the best method when recovering DUD actives, but its performance was similar to that of Xfp. These differences between 2D and 3D fingerprints might be due to the fact that molecular shape is perceived differently when using topological distances between atoms, which overestimate the real distances.

A correlation study was conducted to identify the relationship between 2D and 3D representations [69]. Such study revealed that 3D shapes are more sensitive to shape and flexibility changes than 2D ones, and that they reflect structural features (e.g., conformational flexibility), which strongly depend on the structure. The combination of both 2D and 3D depictions could be useful if the 2D screening is used first. When an alternative technique, pipelining SBVS combined with LBVS, was tested against a typical SBVS method [47], both experiments showed a similar accuracy because the combined method performed better only for 62% of the targets. As it is focused on small fragments rather than large ligands, the computation time employed to assess similarity can be dramatically shortened and, consequently, several screening campaigns can be carried out in a small-medium-sized cluster per day, which could increase the likelihood of finding interesting candidates. This pipeline might improve the traditional SBVS results when a careful selection of fragments is carried out; and, furthermore, it could improve the accuracy of typical LBVS campaigns because structure-related features are considered in the SBVS step. PharmShapeCC was tested in three campaigns of prospective VS [50] and proved to be a convenient method when the goal was to screen large combinatorial libraries, because it is focused on entire libraries rather than individual compounds. Avalanche was also used in two searches for discovering novel compounds, needing <7 min to screen 1.5 million molecules [51]. According to the data, the alignment of the 10000 first hits took longer than the histogram-based comparison, which suggests that histograms can be an interesting choice for comparing molecules. Finally, the machine-learning approach reached a precision of 75% when identifying new Cathepsin-L inhibitors in a database of 825 molecules but, contrary to the other tools, the computation time was very long (1.5 days on one processor) [52].

#### Web servers

Two major aspects should be considered in any evaluation of LBVS servers: performance and usability. The former encompasses the reliability of the results and the time invested to obtain them, while the latter concerns by the parameters provided to setup experiments. Establishing a fair comparison of the web servers in terms of performance and reliability is a difficult task because they should be compared by screening the same query against the same dataset. However, each server screens a different collection of fixed datasets and only a few allow screening a custom library. With regard to speed of calculation, a reasonable comparison is even trickier to achieve, because aspects such as network traffic or server workload cannot be foreseen. For these reasons, a comparison based on accepted metrics such as area under the receiver operating characteristic (AUC-ROC) curve is not possible in this case, and our study will focus on what each server is capable of doing. However, we provide a rough comparison of the number of compounds screened per second (Figures 4–6) and strongly recommend looking up the original publication of each server to obtain an overall picture as regard to their performance. Section 4.1 also gives an estimation of the performance of the underlying similarity algorithms supported by the reported servers.

As regard to the user experience, even though most of the listed servers allow the similarity algorithm to be chosen and some can handle multiple ones (e.g., USR-VS, HybridSim-VS and Superimposé), none of them combines two or more algorithms in the same experiment. Although consensus scoring functions are widely used in molecular docking [119,120] and consensus queries are often present in pharmacophore searches [109,121], only SwissSimilarity and HybridSim-VS present a combined score to combine the results of 2D and 3D screening. However, data fusion could mitigate the dependency on a single type of screening. Among the similarity methods provided, 2D fingerprints are a commonly used choice in most recent servers (ChemMapper, SEABED, SwissSimilarity and HybridSim-VS) due to their low computational cost and the reasonable accuracy of their predictions. The 3D fingerprints are barely accepted and are restricted to the smallest datasets. Of importance, too, is the use of USR in many servers (USR-VS, UFSRAT and ChemMapper). This algorithm owes its success to the simplicity with which it represents molecules. Current servers restrict the use of similarity algorithms to a single one per task. The

Review Banegas-Luna, Cerón-Carrasco &amp; Pérez-Sánchez

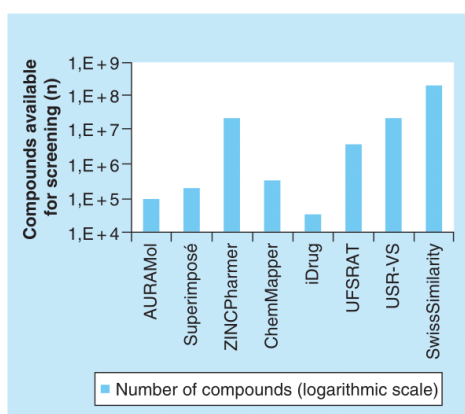


Figure 4. Number of screenable compounds provided by ligand-based virtual screening server.

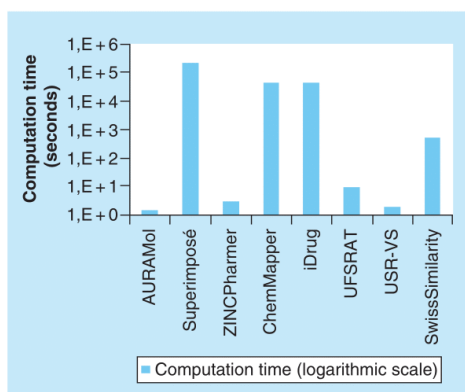


Figure 5. Time in seconds employed by ligand-based virtual screening servers to carry out screening.

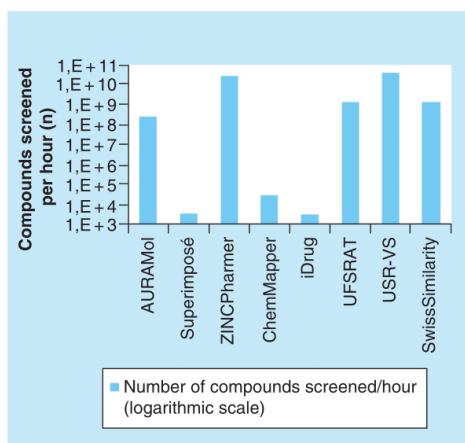


Figure 6. Number of compounds screened per hour by ligand-based virtual screening servers.

libraries to be screened suffer the same restriction, and the screening of multiple databases in the same task is not supported by any server. As an example, ZINCPharmer and USR-VS focus exclusively on the ZINC database. In contrast, the rest of the servers are able to screen many datasets, but only one at a time. It seems logical to think that increasing the amount of screenable compounds should result in better chances of finding successful drug candidates. However, the current trend is to screen small curated libraries for more specific cases, which is considered a cost-effective strategy [122–124]. On the other hand, the screening of custom libraries is an increasingly common feature in the recent servers. iDrug, wwLigSCRre, SEABED and LBVS support such functionality, which might be helpful when the user already has a preprocessed library of ligands to screen.

Speed in delivering results is another important factor in LBVS servers, and can usually be achieved by using more technical approaches. Multithreading is a common technique to screen large datasets by dividing them into small blocks, which are then computed in parallel. Multi-threading is present both in the oldest (e.g., AURAmol) and the newest (e.g., USR-VS and SwissSimilarity) tools. USR-VS also loads all the conformers in the memory to take advantage of the speed of the current hardware. An alternative approach is to improve the performance through the database. This is the case of ZINCPharmer, which uses storage indexes to quickly find the required information. Although the datasets offered by servers contain thousands or millions of compounds, users are frequently interested in the best-ranked ones, so that excluding the least-scored candidates is another common feature in the newest servers. The availability of cutoffs is not a decisive factor but, correctly used, might improve the quality of the results by discarding several unsatisfactory ones.

The continuous progress in the field of LBVS suggests that upcoming web servers should be flexible and scalable enough to absorb such changes. These two features define the architecture of the server, so that its design becomes a crucial point. We have noticed that only ChemMapper clearly defines a component-based architecture. The fact that ChemMapper is the server with the largest number of datasets that can be screened and the highest number of similarity algorithms available, points to good correlation between modularity and flexibility.

### Conclusion & outlook

We have presented a collection of web servers for performing LBVS tasks. Although they all aim to predict the most promising set of drug-like candidates, they employ different parameters to do so. Such parameters are usually very standard, but two key ones can be identified: the similarity algorithm(s) and the total group of dataset(s) to screen. On the basis of these parameters, the typical scenario is the screening of a small curated library by means of a single similarity algorithm. The combination of techniques of different nature is another emerging topic. SBVS and LBVS have already been successfully combined with promising results. Additionally, machine-learning approaches are gaining importance due to their ability to learn from the overwhelming amount of data available. All these novel strategies suggest that the precision of predictions will increase rapidly in forthcoming years; however, to date, they have not shown the desired performance yet, which remains a major challenge for LBVS. To achieve this level of performance, LBVS servers might profit from the continuous advances being made in hardware, software and high-performance computing (HPC) technologies. The vast improvement in computing power made available by recent developments in the area of HPC might help, as well, in the enhancement and refinement of current methodologies and models used routinely nowadays in most VS methods. This can be achieved for instance by adding new terms in the scoring functions, new and more complex descriptors, novel optimization methods, more computationally efficient screening strategies – among others.

### Future perspective

LBVS has grown in importance in recent decades due to its ability to find lead compounds and scaffolds that limit the number of compounds available for experimental testing. However, in the era of big data, a growth in chemical databases, both in terms of size and diversity of the information is expected. Therefore, LBVS web servers will be needed to make an effort to improve on, or at least keep with, current accuracy and performance levels. Future similarity algorithms may not be accurate enough in themselves to reach the precision in predictions. Nevertheless, the combination of methods of a different nature, such as the aforementioned 2D/3D and SBVS/LBVS approaches, is expected to gain importance in coming years because they have already been demonstrated to be successful in previous works. In addition, machine-learning techniques may also gain importance for VS, given their ability to extract knowledge from data, principally due to the expected growth of the chemical databases. In terms of performance, the expected increase in the size of the chemical space will require more computational power to

Review Banegas-Luna, Cerón-Carrasco & Pérez-Sánchez

support the assessment of more compounds and the development of more complex algorithms. In this sense, HPC infrastructures will surely play a decisive role in improving the screening process.

#### Executive summary

##### Molecular representations

- The choice of 2D or 3D representations is a crucial decision in ligand-based virtual screening. The 2D-based algorithms are usually less accurate but faster than 3D based. However, this affirmation strongly depends on the case in question. The selection of the software packages used for each task (e.g., similarity calculation, descriptor extraction and conformer generation) is also dependent on each specific case.
- Publicly accessible chemical databases store an overwhelming amount of data of different natures.
- As regard to dataset size, the current trend is to extract and prepare small curated libraries for screening from the large chemical databases.

##### Web servers

- Web servers for ligand-based virtual screening provide many configuration options to customize tasks.
- Many combinations of parameters remain untested, including the use of multiple similarity searching algorithms for the same task and the screening of customized libraries.
- Web servers tend to restrict the combination of their parameters to preserve the responsiveness of the service.
- The development of high-performance computing techniques (e.g., multi-threading and computing clusters) can help web servers to make better predictions and deliver their results faster.
- The growing number of compounds available in the chemical databases and the diversity of software tools related with virtual screening represent a challenge for both existing and future servers.

#### Conflicts of interest

The authors declare no conflict of interest.

#### Financial & competing interest disclosure

This work has been funded by a grant from the Spanish Ministry of Economy and Competitiveness (CTQ2017-87974-R). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

#### References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Zhang W, Ji L, Chen Y *et al.* When drug discovery meets web search: learning to rank for ligand-based virtual screening. *J. Cheminform.* 13, doi:10.1186/s13321-015-0052-z (2015) (Epub ahead of print).
2. Drews J. Drug discovery: a historical perspective. *Science.* 287(5460), 1960–1965 (2000).
  - A review of drug discovery history, from the early years until high-throughput screening times.
3. Bajorath J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1(11), 882–894 (2002).
4. Willett P. Special issue: chemoinformatics. *Molecules.* 21(4), 535 (2016).
5. Kar S, Roy K. How far can virtual screening take us in drug discovery? *Expert Opin. Drug Discov.* 8(3), 245–261 (2013).
6. Strynadka NC, Jensen SE, Alzari PM, James MN. A potent new mode of  $\beta$ -lactamase inhibition revealed by the 1.7 Å x-ray crystallographic structure of the TEM-1-BLIP complex. *Nat. Struct. Biol.* 3(3), 290–297 (1996).
7. Latek D, Eknomomiuk D, Kolinski A. Protein structure prediction: combining *de novo* modeling with sparse experimental data. *J. Comput. Chem.* 28(10), 1668–1676 (2007).
8. Thompson JM, Sgourakis NG, Liu G *et al.* Accurate protein structure modeling using sparse NMR data and homologous structure information. *Proc. Natl Acad. Sci. USA* 109(25), 9875–9880 (2012).
9. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29, 291–325 (2000).
10. Dalton JAR, Jackson RM. Homology-modelling protein–ligand interactions: allowing for ligand-induced conformational change. *J. Mol. Biol.* 399(4), 645–661 (2010).
11. Leelananda SP, Lindert S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* 12, 2694–2718 (2016).
  - An overview of computational methods used in different aspects of drug discovery. Structure- and ligand-based approaches are discussed.



12. Shanmugam G, Jeon J. Computer-aided drug discovery in plant pathology. *Plant Pathol. J.* 33(6), 529–542 (2017).
13. Glaab E. Building a virtual ligand screening pipeline using free software: a survey. *Brief. Bioinform.* 17(2), 352–366 (2016).
14. Forli S. Charting a path to success in virtual screening. *Molecules* 20(10), 18732–18758 (2015).
15. Danishuddin M, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov. Today* 21(8), 1291–1302 (2016).
16. Shin W-H, Zhu X, Bures M, Kihara D. Three-dimensional compound comparison methods and their application in drug discovery. *Molecules* 20(7), 12841–12862 (2015).
- Reviews some ligand 3D shape comparison methods, including a benchmark of the most representative ones. It also explains the main differences between 1D, 2D and 3D methods.
17. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. Conformer generation with OMEGA: algorithm and validation using high quality structures from the PDB and Cambridge Structural Database. *J. Chem. Inf. Model.* 50(4), 572–584 (2010).
18. Organic chemistry portal. <http://www.organic-chemistry.org>
19. Molecular Discovery MoKa. [www.moldiscovery.com/software/moka/](http://www.moldiscovery.com/software/moka/)
20. Milletti F, Storchi L, Sforza G, Cruciani G. New and original pKa prediction method using grid molecular interaction fields. *J. Chem. Inf. Model.* 47(6), 2172–2181 (2007).
21. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 52(7), 1757–1768 (2012).
22. Kim S, Thiessen PA, Bolton EE *et al.* PubChem substance and compound databases. *Nucleic Acids Res.* 4(44), 1202–1213 (2016).
23. Bento AP, Gaulton A, Hersey A *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42(Database issue), 1083–1090 (2014).
24. Wishart DS, Feunang YD, Guo AC *et al.* Drug Bank 5.0: a major update to the Drug Bank database for 2018. *Nucleic Acids Res.* 4(46), 1074–1082 (2018).
25. Riddigkeit L, van Deursen R, Blum LC, Reymond J-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52(11), 2864–2875 (2012).
26. Pence HE, Williams A. ChemSpider: an online chemical information resource. *J. Chem. Educ.* 87, 1123–1124 (2010).
27. Girke T, Cheng L-C, Raikhel N. ChemMine. A compound mining database for chemical genomics. *Plant Physiol.* 138(2), 573–577 (2005).
28. Seiler KP, George GA, Happ MP *et al.* ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* 36(Database issue), 351–359 (2008).
29. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 4(44), D1045–D1053 (2016).
30. Berman HM, Westbrook J, Feng Z *et al.* The PDB. *Nucleic Acids Res.* 28(1), 235–242 (2000).
31. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28(1), 27–30 (2000).
32. Laney D. 3D Data management: controlling data volume, velocity and variety. *Application Delivery Strategies*(2001). <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
33. Katsila T, Spyroulias GA, Patrinos GP, Matsoukas M-T. Computational approaches in target identification and drug discovery. *Comput. Struct. Biotechnol. J.* 14, 177–184 (2016).
34. Pastur-Romay LA, Cedrón F, Pazos A, Porto-Pazos AB. Deep artificial neural networks and neuromorphic chips for big data analysis: pharmaceutical and bioinformatics applications. *Int. J. Mol. Sci.* 17(1313), 1–26 (2016).
35. Awale M, Visini R, Probst D, Arús-Pous J, Reymond J-L. Chemical space: big data challenge for molecular diversity. *Chimia (Aarau)* 71(10), 661–666 (2017).
36. Karthikeyan M, Vyas R. Role of open source tools and resources in virtual screening for drug discovery. *Comb. Chem. High Throughput Screen.* 18(6), 528–543 (2015).
37. Haga JH, Ichikawa K, Date S. Virtual screening techniques and current computational infrastructures. *Curr. Pharm. Des.* 22(23), 3576–3584 (2016).
- Presents a recent state-of-the-art overview of the technologies applied in virtual screening, and gives a general overview of the virtual screening process.
38. Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. *J. Comput. Chem.* 38(16), 1291–1307 (2017).
39. Krüger J, Thiel P, Merelli I, Grunzke R, Gesing S. Portals and web-based resources for virtual screening. *Curr. Drug Targets* 17(14), 1649–1660 (2016).
- Overview of current online resources related to drug discovery and virtual screening. Topics such as user interfaces and computational performance are also discussed.
40. Chen YC. Beware of docking! *Trends Pharmacol. Sci.* 36(2), 78–95 (2015).

## Review Banegas-Luna, Cerón-Carrasco &amp; Pérez-Sánchez

41. Villoutreix BO, Lagorce D, Labbé CM, Sperandio O, Miteva MA. One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. *Drug Discov. Today* 18(21–22), 1081–1089 (2013).
- A compilation of online tools and databases for drug discovery.
42. Singla D, Dhanda SK, Chauhan JS *et al.* Open source software and web services for designing therapeutic molecules. *Curr. Top. Med. Chem.* 13(10), 1172–1191 (2013).
43. Ramsay RR, Popovic-Nikolic MR, Nikolic K, Uliassi E, Bolognesi ML. A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* 7(1), 3 (2018).
44. Kristensen TG, Nielsen J, Pedersen CNS. Methods for similarity-based virtual screening. *Comput. Struct. Biotechnol. J.* 5(6), 1–6 (2013).
45. Kumar R, Jade D, Gupta D. A novel identification approach for discovery of 5-HydroxyTryptamine 2A antagonists: combination of 2D/3D similarity screening, molecular docking and molecular dynamics. *J. Biomol. Struct. Dyn.* 5, 1–13 (2018).
46. Martínez-Santiago O, Cabrera RM, Marrero-Ponce Y *et al.* Generalized molecular descriptors derived from event-based discrete derivative. *Curr. Pharm. Des.* 22(33), 5095–5113 (2016).
47. Cortés-Cabrera Á, Gago F, Morreale A. A reverse combination of structure-based and ligand-based strategies for virtual screening. *J. Comput. Aided. Mol. Des.* 26(3), 319–327 (2012).
48. Awale M, Jin X, Reymond JL. Stereoselective virtual screening of the ZINC database using atom pair 3D-fingerprints. *J. Cheminform.* 7(3), 1–15 (2015).
49. ElGamacy M, Van Meervelt L. A fast topological analysis algorithm for large-scale similarity evaluations of ligands and binding pockets. *J. Cheminform.* 7(42), doi:10.1186/s13321-015-0091-5 (2015).
50. Muegge I, Zhang Q. 3D virtual screening of large combinatorial spaces. *Methods* 71, 14–20 (2015).
51. Diller DJ, Connell ND, Welsh WJ. Avalanche for shape and feature-based virtual screening with 3D alignment. *J. Comput. Aided. Mol. Des.* 29(11), 1015–1024 (2015).
52. Chen JJF, Visco DP. Developing an *in silico* pipeline for faster drug candidate discovery: virtual high throughput screening with the Signature molecular descriptor using support vector machine models. *Chem. Eng. Sci.* 159, 31–42 (2017).
53. Leach AR, Gillet VJ. *An Introduction to Chemoinformatics (Revised Edition)*. Springer, Dordrecht, The Netherlands (2007).
54. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics (2nd Edition)*. Wiley-VCH, Weinheim, Germany (2009).
55. Li ZR, Han LY, Xue Y *et al.* MODEL – Molecular Descriptor Lab: a web-based server for computing structural and physicochemical features of compounds. *Biotechnol. Bioeng.* 97(2), 389–396 (2007).
56. Cao DS, Xu QS, Hu QN, Liang YZ. ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* 29(8), 1092–1094 (2013).
57. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 31, 2967–2970 (2010).
58. Katritzky AR, Karelson M, Petrukhin R. CODESSA Pro. [www.codessa-pro.com/](http://www.codessa-pro.com/)
59. ADRIANA.Code. <https://www.mn-am.com/products/adriana-code>.
60. Molecular Operating Environment (MOE) 2013.08. Chemical Computing Group Inc., Montreal, QC, Canada. <http://www.chemcomp.com>.
61. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E. The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* 43, 493–500 (2003).
62. Landrum G. RDKit: open-source cheminformatics software (2016). [www.rdkit.org](http://www.rdkit.org).
63. Preeja MP, Palivela H, Soman KP, Kharkar PSK. Ligand-based virtual screening using random walk kernel and empirical filters. *Procedia Comput. Sci.* 57, 418–427 (2015).
64. Kayik G, Tüzün NŞ, Durdagi S. Investigation of PDE5/PDE6 and PDE5/PDE11 selective potent tadalafil-like PDE5 inhibitors using combination of molecular modeling approaches, molecular fingerprint-based virtual screening protocols and structure-based pharmacophore development. *J. Enzyme Inhib. Med. Chem.* 32(1), 311–330 (2017).
65. Fernández-De Gortari E, García-Jacas CR, Martínez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. *J. Cheminform.* 9(1), 1–9 (2017).
66. Sauer WHB, Schwarz MK. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* 43, 987–1003 (2003).
67. Ballester PJ, Richards WG. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* 28(10), 1711–1723 (2007).
68. Schreyer AM, Blundell T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminform.* 4(27), doi: 10.1186/1758-2946-4-27 (2012).
69. Dobi K, Hajdú I, Flachner B *et al.* Combination of 2D/3D ligand-based similarity search in rapid virtual screening from multimillion compound repositories. Selection and biological evaluation of potential PDE4 and PDE5 inhibitors. *Molecules* 19(6), 7008–7039 (2014).

70. Willett P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* 11(23–24), 1046–1053 (2006).
71. Rarey M, Stahl M. Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aided Mol. Des.* 15(6), 497–520 (2001).
72. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. *Science* (80), 221(4612), 709–713 (1983).
73. Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* 8, 6–13 (2017).
74. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief. Bioinform.* 18(5), 851–869 (2017).
75. Malta TM, Sokolov A, Gentles AJ *et al.* Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* 173(2), 338–354 (2018).
76. Ridder L, Wang H, de Vlieg J, Wagener M. Revisiting the rule of five on the basis of pharmacokinetic data from rat. *ChemMedChem.* 6(11), 1967–1970 (2011).
77. Zolfaghari N. Molecular docking analysis of nitisinone with homogentisate 1,2 dioxygenase. *Bioinformation* 13(5), 136–139 (2017).
78. Zolfaghari N. Competitive rational inhibitor design to 4-maleyl-aceto-acetate isomerase. *Bioinformation* 13(5), 140–143 (2017).
79. Ahmad SS, Akhtar S, Rizvi SMD *et al.* Screening and elucidation of selected natural compounds for anti-Alzheimer's potential targeting BACE-1 enzyme: a case computational study. *Curr. Comput. Aided Drug Des.* 13(4), 311–318 (2017).
80. Leeson PD, Davis AM, Steele J. Drug-like properties: guiding principles for design – or chemical prejudice? *Drug Discov. Today Technol.* 1(3), 189–195 (2004).
81. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53(7), 2719–2740 (2010).
82. Veber DF, Johnson SR, Cheng H, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623 (2002).
83. Hughes JD, Blagg J, Price DA *et al.* Physicochemical drug properties associated with *in vivo* toxicological outcomes. *Bioorganic Med. Chem. Lett.* 18(17), 4872–4875 (2008).
84. Rampogu S, Baek A, Zeb A, Lee KW. Exploration for novel inhibitors showing back-to-front approach against VEGFR-2 kinase domain (4AG8) employing molecular docking mechanism and molecular dynamics simulations. *BMC Cancer* 18(1), 1–21 (2018).
85. Alam S, Khan F. Virtual screening, docking, ADMET and system pharmacology studies on garcinia caged xanthone derivatives for anticancer activity. *Sci. Rep.* 8(1), 5524 (2018) (Epub ahead of print).
86. Mansuri R, Kumar A, Rana S *et al.* *In vitro* evaluation of antileishmanial activity of computationally screened compounds against ascorbate peroxidase to combat amphotericin B drug resistance. *Antimicrob. Agents Chemother.* 61(7), 1–25 (2017).
87. Kong X, Quin J, Li Z *et al.* Development of a novel class of B-Raf(V600E)-selective inhibitors through virtual screening and hierarchical hit optimization. *Org. Biomol. Chem.* 10(36), 7402–7417 (2012).
88. York TU of. AURAMol. <https://www.cs.york.ac.uk/auramol/>
89. Klinger S, Austin J. Chemical similarity searching using a neural graph matcher. *13th European Symposium on Artificial Neural Networks*. Bruges, Belgium (2005). <https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2005-91.pdf>
90. Klinger S, Austin J. A neural supergraph matching architecture. *Proc. Int. Jt. Conf. Neural Networks* 1–5, 2453–2458 (2005).
91. Bauer RA, Bourne PE, Formella A *et al.* Superimposé: a 3D structural superposition server. *Nucleic Acids Res.* 36(Web server issue), W47–W54 (2008).
92. Sperandio O, Petitjean M, Tuffery P. wwLigCSRre: a 3D ligand-based server for hit identification and optimization. *Nucleic Acids Res.* 37(Suppl. 2), 504–509 (2009).
93. Koes DR, Camacho CJ. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.* 40(Web server issue), W409–W414 (2012).
94. Gong J, Cai C, Liu X *et al.* ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics* 29(14), 1827–1829 (2013).
95. Wang X, Chen H, Yang F *et al.* iDrug: a web-accessible and interactive drug discovery and design platform. *J. Cheminform.* 6(28), doi: 10.1186/1758-2946-6-28 (2014).
96. Zheng M, Liu Z, Yan X, Ding Q, Gu Q, Xu J. LBVS: an online platform for ligand-based virtual screening using publicly accessible databases. *Mol. Divers.* 18(4), 829–840 (2014).
97. Shave S, Blackburn EA, Adie J *et al.* UFSRAT: Ultra-Fast Shape Recognition with Atom Types – the discovery of novel bioactive small molecular scaffolds for FKBP12 and 11βHSD1. *PLoS ONE* 10(2), 1–15 (2015).
98. Plos T, Staff ONE. Correction: UFSRAT: Ultra-Fast Shape Recognition with Atom Types – the discovery of novel bioactive small molecular scaffolds for FKBP12 and 11βHSD1. *PLoS ONE* 10(3), e0122658 (2015).
99. Fenollosa C, Otón M, Andrio P, Cortés J, Orozco M, Goñi JR. SEABED: Small molEcule activity scanner weB servicE baseD. *Bioinformatics* 31(5), 773–775 (2015).

100. Li H, Leung K-S, Wong M-H, Ballester PJ. USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. *Nucleic Acids Res.* 44(W1), W436–441 (2016).
101. Zoete V, Daina A, Bovigny C, Michielin O. SwissSimilarity: a web tool for low to ultra high throughput ligand-based virtual screening. *J. Chem. Inf. Model.* 56(8), 1399–1404 (2016).
102. Shang J, Dai X, Li Y, Pistozzi M, Wang L. HybridSim-VS: a web server for large-scale ligand-based virtual screening using hybrid similarity recognition techniques. *Bioinformatics* 33(21), 3480–3481 (2017).
103. Yu W, MacKerell AD Jr. Computer-aided drug design methods. *Methods Mol. Biol.* 1520, 85–106 (2017).
104. Gomes MN, Muratov EN, Pereira M *et al.* Chalcone derivatives: promising starting points for drug design. *Molecules* 22(8), 1210 (2017).
105. Armstrong MS, Morris GM, Finn PW, Sharma R, Moretti L, Cooper RI. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided Mol. Des.* 24(9), 789–801 (2010).
106. Silicos IT (2014). <http://openbabel.org/docs/dev/Fingerprints/spectrophore.html>
107. Silicos-IT. Silicos-IT Home page (2014). <http://silicos-it.be.s3-website-eu-west-1.amazonaws.com>
108. Krau EF. *Taxicab Geometry (First Edition)*. Addison-Wesley, Menlo Park, CA, USA (1975).
109. Koes DR, Camacho CJ, Pharmar: efficient and exact pharmacophore search. *J. Chem. Inf. Model.* 51(6), 1307–1314 (2011).
110. Sanders MPA, Barbosa JM, Zarzycka B *et al.* Comparative analysis of pharmacophore screening tools. *J. Chem. Inf. Model.* 52(6), 1607–1620 (2012).
111. Liu X, Jiang H, Li H. SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. *J. Chem. Inf. Model.* 51(9), 2372–2385 (2011).
112. Lu W, Liu X, Cao X *et al.* SHAFTS: a hybrid approach for 3D molecular similarity calculation. 2. Prospective case study in the discovery of diverse p90 ribosomal S6 protein kinase 2 inhibitors to suppress cell migration. *J. Med. Chem.* 54(10), 3564–3574 (2011).
113. Zhou W, Liu X, Tu Z *et al.* Discovery of pteridin-7(8H)-one-based irreversible inhibitors targeting the epidermal growth factor receptor (EGFR) kinase T790M/L858R mutant. *J. Med. Chem.* 56(20), 7821–7837 (2013).
114. Utepergenov D, Derewenda U, Olekhnovich N *et al.* Insights into the inhibition of the p90 ribosomal S6 kinase (RSK) by the flavonol glycoside SL0101 from the 1.5Å crystal structure of the N-terminal domain of RSK2 with bound inhibitor. *Biochemistry* 51(33), 6499–6510 (2012).
115. Zhang S, Shan L, Li Q *et al.* Systematic analysis of the multiple bioactivities of green tea through a network pharmacology approach. *Evidence-based Complement. Altern. Med.* 2014, 512081 (2014).
116. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42(6), 1273–1280 (2002).
117. Open Eye Scientific. <https://www.eyesopen.com>
118. Haque IS, Pande VS, Walters WP. SIML: a fast SIMD algorithm for calculating LINGO chemical similarities on GPUs and CPUs. *J. Chem. Inf. Model.* 50(4), 560–564 (2010).
119. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* 31(2), 455–461 (2010).
120. Ferreira LG, dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. *Molecules.* 20(7), 13384–13421 (2015).
121. Wei N, Hamza A. SABRE: ligand/structure-based virtual screening approach using consensus molecular-shape pattern recognition. *J. Chem. Inf. Model.* 54, 338–346 (2014).
122. Jacoby E, Davies J, Blommers MJJ. Design of small molecule libraries for NMR screening and other applications in drug discovery. *Curr. Top. Med. Chem.* 3, 11–23 (2003).
123. Miller JL. Recent developments in focused library design: targeting gene families. *Curr. Top. Med. Chem.* 6(1), 19–29 (2006).
124. Mok NY, Brenk R. Mining the ChEMBL database: an efficient chemoinformatics workflow for assembling an ion channel-focused screening library. *J. Chem. Inf. Model.* 51(10), 2449–2454 (2011).

### 3.4 BRUSELAS: HPC GENERIC AND CUSTOMIZABLE SOFTWARE ARCHITECTURE FOR 3D LIGAND-BASED VIRTUAL SCREENING OF LARGE MOLECULAR DATABASES

<b>Título</b>	BRUSELAS: HPC generic and customizable software architecture for 3D ligand-based virtual screening of large molecular databases
<b>Autores</b>	Banegas-Luna AJ, Cerón-Carrasco JP, Pérez-Sánchez H.
<b>Revista</b>	Journal of Chemical Information and Modeling
<b>Año</b>	2019
<b>Volumen</b>	
<b>Páginas</b>	
<b>Estado</b>	Publicado online
<b>DOI</b>	<a href="https://doi.org/10.1021/acs.jcim.9b00279">https://doi.org/10.1021/acs.jcim.9b00279</a>
<b>IF(2017)</b>	3.804
<b>Categoría</b>	Chemistry, Medicinal, 11/59, Q1 Computer Science, Interdisciplinary Applications, 15/105, Q1 Computer Science, Information Systems, 20/148, Q1

#### Contribución del Doctorando

El doctorando Antonio Jesús Banegas Luna declara ser el autor principal y contribuyente principal del artículo *BRUSELAS: HPC generic and customizable software architecture for 3D ligand-based virtual screening of large molecular databases*.

## BRUSELAS: HPC Generic and Customizable Software Architecture for 3D Ligand-Based Virtual Screening of Large Molecular Databases

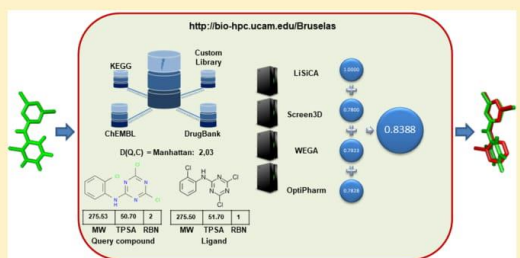
Antonio J. Banegas-Luna,<sup>\*,†</sup> José P. Cerón-Carrasco,<sup>†</sup> Savíns Puertas-Martín,<sup>‡</sup> and Horacio Pérez-Sánchez<sup>\*,†</sup>

<sup>†</sup>Bioinformatics and High Performance Computing Research Group (BIO-HPC), Computer Engineering Department, Universidad Católica San Antonio de Murcia (UCAM), Campus de los Jerónimos s/n, 30107 Murcia, Spain

<sup>‡</sup>Supercomputing - Algorithms Research Group (SAL), Department of Informatics, University of Almería, Agrifood Campus of International Excellence, ceiA3, Almería, 04120, Spain

### Supporting Information

**ABSTRACT:** BRUSELAS (balanced rapid and unrestricted server for extensive ligand-aimed screening) is a novel, highly efficient web software architecture for 3D shape and pharmacophore searches in *off the cuff* libraries. A wide panel of shape and pharmacophore similarity algorithms are combined to avoid unbiased results while yielding consensus scoring functions. To evaluate its reliability, BRUSELAS was tested against other similar servers (e.g., USR-VS, SwissSimilarity, ChemMapper) to search for potential antidiabetic drugs. A web tool is developed for users to customize their tasks and is accessible free of any charge or login at <http://bio-hpc.eu/software/Bruselas>. Source code is available on request.



## 1. INTRODUCTION

Virtual screening (VS) is a set of *in silico* techniques that aims to increase the chances of finding novel hit and lead compounds. Despite its moderate computational cost, VS remains one of the preferred options for assessing large chemical spaces and for reducing the time spent on *in vivo* experimentation.<sup>1</sup> This approach has gained importance during recent decades for studying of a variety of diseases such as diabetes, malaria, and Parkinson's disease.<sup>2–12</sup> The wide diversity of VS techniques may be classified into two major groups, depending on whether the 3D structure of the target is known (structure-based, SBVS) or not (ligand-based, LBVS). LBVS relies on a knowledge of small molecules that bind the target of interest and encompasses a collection of methods, including similarity searching, pharmacophore modeling, and QSAR (quantitative structure activity relationship).<sup>13</sup> In the search for a cost-effective solution, similarity searching is the favorite approach, whenever possible, due to its simplicity.<sup>14</sup> The underlying idea behind similarity searching is to build a fingerprint for molecules that can be easily compared, while simultaneously maintaining the information necessary to identify similar biological activity. An alternative to this approach is pharmacophore modeling, which aligns two or more molecules to identify the shared pharmacophore features between them. In this case, molecular similarity is represented in terms of compounds whose pharmacophoric features are quite similar to the pattern identified.<sup>15</sup>

To provide computational services accessible to any users, several VS web servers have emerged concomitant with the development of web technologies. Usually, online VS servers focus on the application of one or many techniques and allow users to set a collection of parameters, including the databases to be screened and the similarity algorithms to be applied.<sup>16</sup> Although VS is widely used, some critical issues remain unsolved. For example, millions of compounds have been studied and their details published in the literature, but only those fulfilling drug-related requirements are of interest. In addition, computational resources may also represent a bottleneck when the chemical space is very large or the tasks to perform are computationally expensive.<sup>17</sup> Finally, the large number of existing compounds, tools, and techniques complicates the development of research on VS.

To help to solve such critical issues, we have developed BRUSELAS (balanced rapid and unrestricted server for extensive ligand-aimed screening), which is a software architecture focused on 3D LBVS. BRUSELAS integrates several software tools (e.g., Obabel, WEGA, DRAGON) and hides their details from the final user (Table 1). As an additional feature, it maintains a large database of compounds imported from publicly accessible databases and curated for their use in VS. That database allows the server to suggest a suitable library containing the most promising compounds for

Received: April 2, 2019

Published: May 10, 2019

**Table 1. List of Software Tools Embedded in BRUSELAS**

task	tool	references
calculation of descriptors	DRAGON	18
chemical databases	DrugBank, ChEMBL, KEGG, DIA-DB	19–22
remove salts from conformers	Standardizer	23
calculation of conformers	Omega2	24
shape similarity	WEGA, LiSiCA, Screen3D, OptiPharm	25–28
pharmacophore modeling	SHAFTS	29
molecular conversion	Obabel, Molconvert	23 and 30
3D visualization	Jmol	31

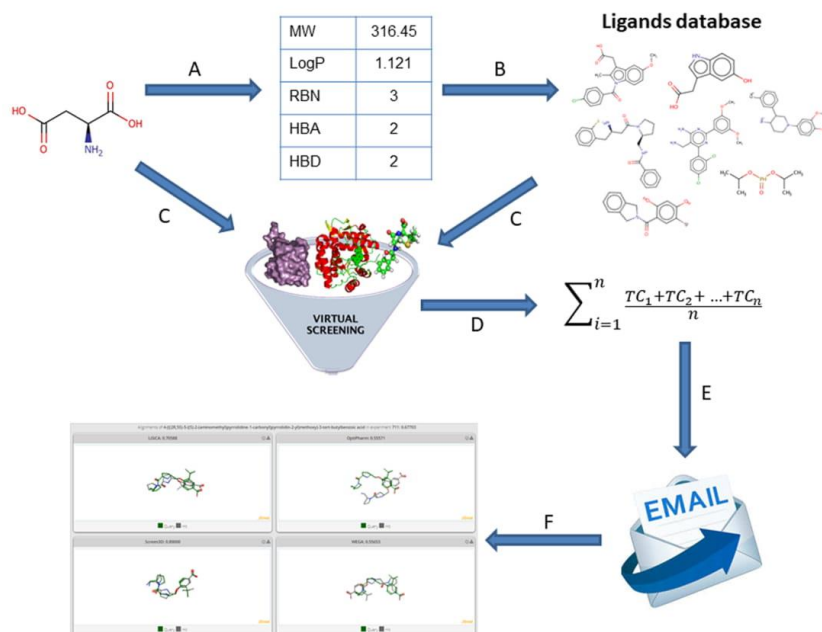
each given query molecule. The screening phase is complemented with a collection of statistical functions to make predictions more effective. Since efficiency is a crucial point in VS, BRUSELAS profits from a high performance computing (HPC) platform for delivering results with a delay of a few minutes only. To ease the use of the architecture and the interpretation of its predictions, a web tool has been developed and is available without fee or login at <http://bio-hpc.eu/software/Bruselass>.

The present work aims to help to make VS available to general scientists by merging advanced features in a transparent and friendly web-based environment. It provides a software architecture for the configuration and execution of 3D

virtual screening tasks, which implements a workflow specifically designed for handling a large number of compounds. Such an architecture can be potentially used by researchers in a number of areas related to drug discovery. On the other hand, it opens up a way to investigate these areas of research, using highly customizable tools in web environments, which can take advantage of the most recent computational advances, e.g., distributed computing, to predict those compounds with the greatest possibilities of becoming drugs in the future. This manuscript is structured as follows: Section 2 introduces an overview of BRUSELAS architecture, the underlying workflow implemented and the main tools it uses. Section 3 assesses the use of the web tool and highlights its main features and the options available. In an attempt to provide a fair but critical assessment of the developed tool, Section 4 compares the reliability and the accuracy of the architecture with other similar servers by using two well-known drugs as queries. The main conclusions reached concerning BRUSELAS in its current version as well as forthcoming research lines are outlined.

## 2. MATERIALS AND METHODS

**2.1. Screening Workflow.** BRUSELAS implements the workflow illustrated in Figure 1. The protocol starts when a user requests a new screening task from the web tool. Although several parameters are available for customizing the tasks, only three of them are mandatory: the screening algorithm(s), the database(s) to be screened and the e-mail address to which the



**Figure 1.** Graphic representation of BRUSELAS workflow for virtual screening tasks. (A) If the server needs to build a library, it calculates the selected descriptors from the query. (B) The calculated descriptors are used for selecting the most promising compounds from the database. (C) Once the library is calculated, both query and library are supplied for 3D screening. (D) If needed, the individual scores are combined in a global score. (E) Users are notified by e-mail when the results are available on the website. (F) Users can analyze the results on the web tool.

results will be sent. Users can choose to submit their own libraries or ask the server to build one *ad hoc* for the given query. If the users decide to screen the BRUSELAS libraries, they can customize the selection of the compounds that will be included in their library. Among the parameters of this step is the set of molecular descriptors that will be used, along with a distance function (e.g., Manhattan), to find the most promising compounds for the 3D screening stage. The descriptors of the ligands are already calculated with the DRAGON<sup>18</sup> tool and stored in a relational database to save time, whereas the ones from the query are calculated every time, unless the users upload their own library. DRAGON is a powerful tool which is able to calculate almost 4900 descriptors. Since this is a very large number of descriptors, the server makes an initial proposal with the most common ones (molecular weight, number of rotatable bonds, number of hydrogen acceptor/donor atoms, topological polar surface area, and octanol–water partition coefficient), but the users have the freedom to modify the proposal by selecting those that best suit their needs (Figure 1A). Then, a fingerprint is built with the selected descriptors, which is used to reduce the number of ligands that will continue to the 3D screening stage (Figure 1B). Next, both the query and the library are sent to a supercomputing cluster for detailed 3D ligand-based screening (Figure 1C). In cases where many similarity algorithms are involved, the server has to deal with several scores arising from each of them. This situation is handled by applying a consensus scoring function, which combines the individual scores into a single one for each compound (Figure 1D). Once the screening is done, all the output generated by the algorithms is interpreted, adapted to BRUSELAS, and stored in the relational database. Finally, the users are notified by e-mail that their results are available on the website (Figure 1E).

**2.2. Compound Database.** As mentioned above, BRUSELAS may have to propose a suitable library for screening when the user does not supply one. For this reason, the server maintains a database of compounds which are imported from the Food and Drug Administration (FDA) data set deposited in DrugBank<sup>19</sup> (1760 compounds), the set of active compounds in ChEMBL 21<sup>20</sup> (1 578 131), KEGG<sup>21</sup> Compounds and Drugs subsets (23 667) and an *in-house* library of antidiabetics, DIA-DB<sup>22</sup> (186). Although a 2D representation of such compounds can be freely obtained, they have been curated for VS in a preparation protocol because most of the similarity algorithms supported by BRUSELAS require many conformations to cover the chemical space. Accordingly, the first step of the protocol is to obtain the 3D representation of all the compounds to better handle flexibility. To find the 3D conformer that is most likely to occur in Nature, a Merck molecular force field (MMFF94) is applied since it has been demonstrated to be suitable for small organic molecules.<sup>32</sup> Next, salts are removed with the standardizer<sup>23</sup> tool, and up to 10 conformers are calculated with the omega2<sup>24</sup> tool for those compounds having between 1 and 10 rotatable bonds. In this step, we set a root-mean-square distance (RMSD) of 0.7 to ensure a diversity of conformers and leave the rest of parameters at their default values. Conformer generation is repeated with less restrictive conditions for those compounds that do not generate any conformer previously. Finally, those compounds containing a heavy metal (e.g., chromium) or a boron atom bonded to a halogen that are not supported by the force field MMFF94 are not imported. The curation process results in a data set of 7 473 006 conformers

available for screening. This number represents an average of eight conformers for compounds with up to 10 rotors. Additionally, a collection of key terms, including bibliography, descriptions, and related targets, is extracted from the source databases and imported into BRUSELAS. Such information may be used for the creation of suitable libraries when users want to narrow the scope of the job to a specific disease or target.

**2.3. Methods Available for 3D Virtual Screening.** BRUSELAS supports both similarity and pharmacophore searching, both of which can be launched from the same interface. In order to obtain the most accurate predictions, BRUSELAS does not rely on one single similarity algorithm, but it proposes four 3D algorithms, each of a different nature, which can be combined in the same task to avoid biased results. The available algorithms are WEGA,<sup>25</sup> LiSiCA,<sup>26</sup> Screen3D,<sup>27</sup> and OptiPharm.<sup>28</sup> The server takes care of the format conversions to make each algorithm work with the required molecular format. Such algorithms represent four different complementary approaches for calculating similarity. Whereas WEGA evaluates similarity by means of weighted atomic Gaussian functions, LiSiCA is based on graph theory. On the other hand, Screen3D calculates a similarity score based on the atomic distances between query and ligands. Note that BRUSELAS configures Screen3D to handle both the query and the ligand, enabling the molecules to be rotated, in order to make more accurate estimations. The last approach, OptiPharm, is a parallelized evolutionary algorithm that needs large populations to analyze the search space and find accurate solutions.

Pharmacophore modeling is performed by SHAFTS,<sup>29</sup> which is a pharmacophore matching method based on a feature triplet hashing and searching algorithm. SHAFTS is suitable for large-scale VS with single or multiple bioactive compounds, and it has been successfully used in many studies, such as the search of p90 ribosomal S6 inhibitors and the development of B-Raf(V600E) inhibitors.<sup>33,34</sup>

**2.4. Consensus Scoring Functions.** LBVS predictions may be influenced by the algorithm used to assess similarity, leading to biased results. This issue might be addressed by combining multiple similarity algorithms, but it requires the fusion of individual scores. Although consensus scoring functions are frequently used in SBVS,<sup>35</sup> they are rarely used in LBVS because ligand-based servers do not allow the combination of different algorithms. However, BRUSELAS does not restrict the selection of similarity algorithms to one single task; the output provides three consensus functions: the mean, the weighted mean, and the maximum score. Whereas the weighted mean is used to assign a customized rate to each algorithm and is especially useful to return unbiased results by a concrete algorithm, the mean function rates every algorithm with the same weight. Furthermore, the maximum score function returns the best score assigned by any of the algorithms applied. This set of functions covers all the possible approaches, from the most optimistic with the maximum score to the most conservative with the mean score.

**2.5. Computation Time.** The screening of large data sets is a time-consuming process whose duration usually grows with the size of the data set. HPC techniques are typically applied to speed up that task, and they have gained importance in the context of VS during recent years.<sup>36</sup> BRUSELAS also profits from HPC by running similarity calculations on a supercomputing cluster. With this approach, the screenable



**A**

Type of calculation **Search** Library building Score Submit

**Algorithm:**

LISICA  OptiPharm  Screen3D  WEGA

**Select query molecule:**

Upload Ligand  
 Ligand (Smiles)  
 Draw Ligand

**Select target library (optional):**

BRUSELAS libraries  
 DIA-DB  DrugBank  ChEMBL  KEGG  
 Custom library  
 Ningún archivo seleccionado

**B**

Type of calculation Search **Library building** Score Submit

**Descriptor extraction software:**

DRAGON

**Select descriptors (optional):** A number between 1 and 100 descriptors must be selected.

Ghose-Crippen octanol-water partition coeff. (logP)  molecular weight  number of acceptor atoms for H-bonds (N,O,F)  
 number of donor atoms for H-bonds (N and O)  number of rotatable bonds  
 topological polar surface area using N,O,S,P polar contributions

**Distance function:**

Manhattan distance  Euclidean distance  Range values

**Compounds related with the following terms:**

**C**

Type of calculation Search Library building **Score** Submit

**Consensus scoring function (optional):**

Arithmetic mean  Weighted arithmetic mean  Maximum value

**Similarity cutoff [0,1] (optional):**

**Maximum number of hits (optional):**

**D**

Type of calculation Search Library building Score **Submit**

**Job description (optional):**

**Email:**

**SEND EXPERIMENT**

**Figure 2.** Task configuration screen of BRUSELAS. (A) Selection of similarity algorithms, query and databases. (B) Users can configure the creation of an *off the cuff* library for screening. (C) Configuration of results and consensus scoring function. (D) A short text can be included in the notification e-mail.

**Parameters of experiment 906**

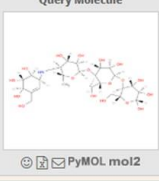
Submitted by	alpanegas@alu.ucom.edu	Start/End date	30/03/2019 22:08:10 - 30/03/2019 22:25:45
Type of calculation	Similarity searching	Similarity software	WEGA, Screen3D, OptiPharm
Scoring cutoff	0	No. Results requested	100
Distance function	Manhattan distance	Consensus function	AVG
Description	acarbose + wega + screen3d + optipharm + drugbank		

[View more](#)

Show  entries Search:  Extended view

Ranking	Compound	Score[0,1]	Alignment
1	Acarbose	0.71856	
2	Ergotamine	0.62548	
3	Cefmetazole	0.59978	
4	Cefazolin	0.58376	
5	Deserpidine	0.55670	
6	Ertapenem	0.55237	
7	Dihydroergotamine	0.55136	
8	Piperacillin	0.54559	
9	Clonidine	0.54555	

**Query Molecule**



[PyMOL mol2](#)

**References**

[1] [3] [4] [5] [6] [7] [9] [10] [12] [13]

**Figure 3.** Analysis of results screen in BRUSELAS.

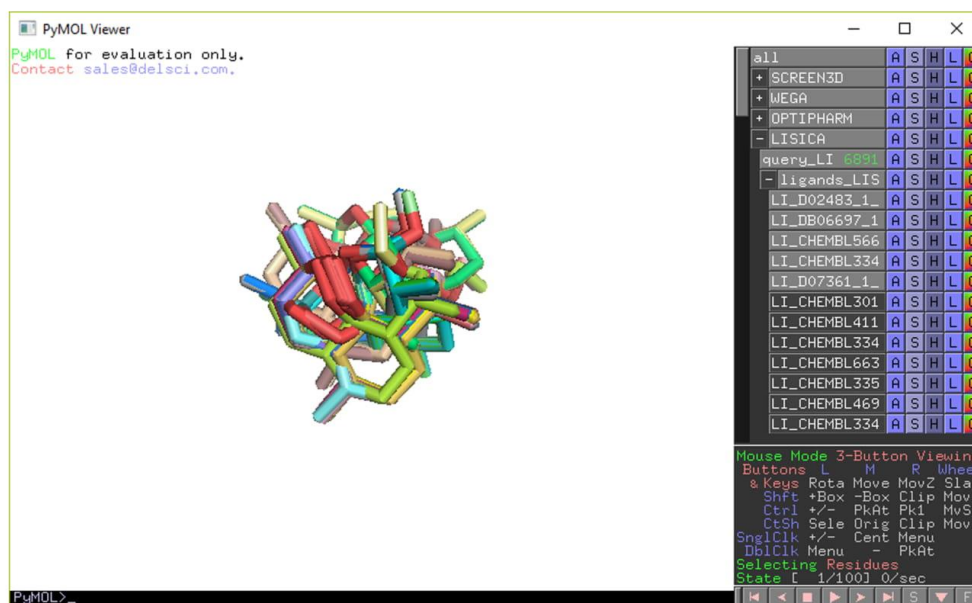


Figure 4. PyMOL session downloaded from BRUSELAS.

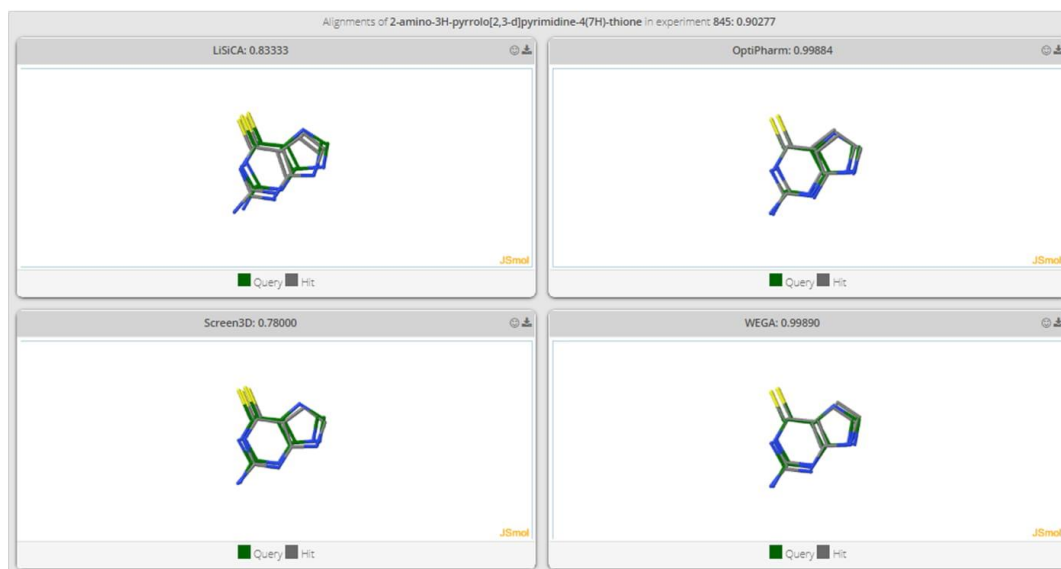


Figure 5. Alignment of query and ligand generated by all similarity algorithms in BRUSELAS.

compounds are grouped into small blocks, and then every group is compared against the query in parallel. As the server is currently configured, a standard task using the default parameters and involving the four algorithms available will deliver the results in less than 1 h. It should be noticed that, although the 100 best-ranked hits are returned by default, the library suggested by the server is larger. By using HPC and

limiting the number of conversions among formats, the results are rapidly delivered.

### 3. WEB TOOL USAGE

**3.1. Input Data.** Before starting the screening process, BRUSELAS needs to collect input parameters, which are gathered from the web tool. Although only a few fields are

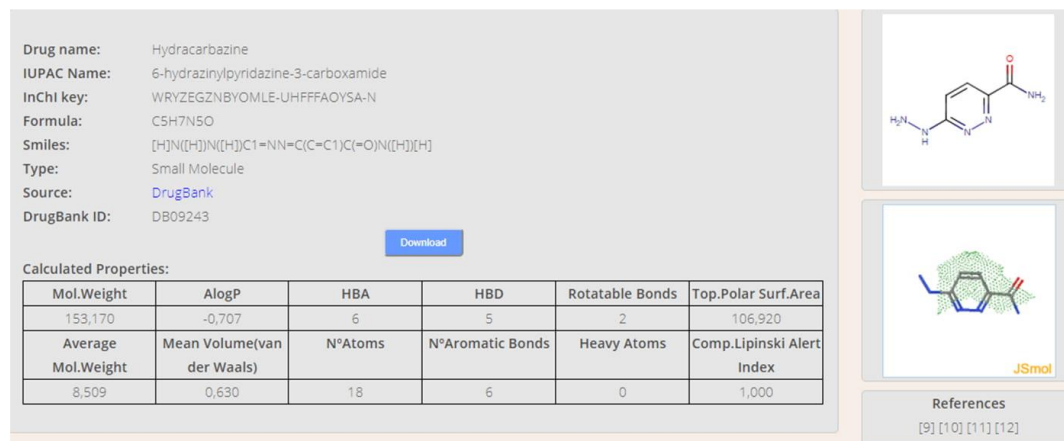


Figure 6. Detailed information on a compound selected from BRUSELAS' database.

required (query, library, similarity algorithm, and e-mail), the server accepts many other options to customize tasks, including the maximum number of hits to return and the scoring cutoff. Input parameters are organized in four tabs according to their purposes, as shown in Figure 2.

One of the most interesting abilities of BRUSELAS is the combination of many similarity algorithms in the same task and the screening of a custom database. Since the server relies on Obabel<sup>30</sup> and Molconvert<sup>23</sup> to carry out conversions, the most common formats are supported, including sdf, mol2, and pdb. Another distinguishing feature is the ability to create an *off the cuff* library on the basis of some key terms. When any keyword is introduced, the server searches for those compounds linked to the given terms and creates a library with them. This novel feature is useful for creating specific libraries for a given disease or focusing searches on a certain target. In addition, tasks can also be labeled with a short text, which is included in the notification e-mail, to differentiate them.<sup>37</sup>

**3.2. Analysis of Results.** Users are informed by e-mail as soon as the calculation is completed.<sup>37</sup> The notification includes a link to the result explorer screen, which is organized in three blocks: the summary header, the query molecule and the list of hits (Figure 3). The header shows a summary of the input parameters used to set the task up. On the right, a 2D representation of the query is displayed. The user can manipulate the query and the results with the four buttons available under the query providing the following functionalities: (1) download the SMILES representation of the query; (2) download compound names and scores in an Excel file; (3) share the results with other users by e-mail; (4) download the hits in a PyMOL session where the five best-ranked compounds are selected by default (Figure 4); (5) download all the aligned ligands and the query in separate mol2 files for their offline manipulation. The third block lists candidate compounds sorted by score, which is expressed in terms of the Tanimoto coefficient, in descending order. Although this score is the final one calculated by BRUSELAS, individual scores predicted by each algorithm are also available in the same table by enabling the "Extended view" check. Next to the scores, the "Alignment" button opens a new window where both query

and ligand are conveniently aligned in the Jmol<sup>31</sup> applet (Figure 5). This feature helps users to visualize, at a glance, how both molecules are aligned by each algorithm, with the possibility of downloading each aligned ligand individually.

**3.3. Compound Explorer.** The compound database maintained by BRUSELAS is available not only for screening but also for exploring its main properties. Compounds can be filtered by a variety of fields, including chemical name, International Union of Pure and Applied Chemistry (IUPAC) name, and SMILES representation. To obtain further information about a molecule, users just click on the name and they are directed to a detail screen (Figure 6), where additional information is displayed (e.g., typical molecular descriptors, link to the original source).

## 4. CASE STUDIES

**4.1. Search for Malaria Drugs.** Artemether is a drug frequently used in treatments against Malaria since its first approval in 2009, especially in Africa and Asia, either alone or in combination with other drugs such as Lumefantrine or Amodiaquine.<sup>38–41</sup> With the aim of testing the reliability of BRUSELAS in the simplest case, we tested all the combinations of similarity algorithms with a copy of artemether as query. As expected, artemether was predicted as the best-ranked candidate with 100% similarity in all the cases. This result demonstrates that BRUSELAS behaves as expected in a simple case, independent of the combination of algorithms selected.

**4.2. Antidiabetic Compounds.** Type 2 diabetes mellitus (T2DM) is a frequent form of insulin resistance that maintains glucose homeostasis by increasing the release of insulin.<sup>49</sup> Acarbose is a glucosidase inhibitor which has been extensively applied in T2DM cases.<sup>50–53</sup> In our case, acarbose was used as the query to conduct a second experiment with the aim of assessing the reliability of BRUSELAS predictions compared with some similar servers. Every server was supplied a copy of acarbose downloaded from ZINC<sup>54</sup> database, and the FDA approved drugs data set from DrugBank was screened when possible. In cases where acarbose is present in the selected database, it should be returned as the best-ranked hit. Moreover, the expected score should be as close as possible

Table 2. Assessment of the Efficacy of Some Relevant LBVS Servers in the Context of T2DM

VS server	similarity method	library to screen	rank acarbose	score	ref.
USR-VS	USR	ZINC	not found	–	42
	USCAT		not found	–	
SwissSimilarity	FP2	Drugs Approved	1	1.000	43
	ElectroShape 5D		1	0.907	
	Spectrophores		2	0.823	
	Shape-IT		not found	–	
BRUSELAS <sup>a</sup>	Align-IT		not found	–	
	All shape similarity	DrugBank	1	0.757	
SHAFTS			not found	–	
Superimposé	Score1	Superdrug	not found	–	44
	Score1	Ligand Depot	not found	–	
	Sd_best_compare	Superdrug	not finished	–	
HybridSim-VS	FP2 + WEGA	DrugBank Approved	1	0.693	45
	MACCS + WEGA		1	0.685	
ChemMapper	FP2	DrugBank Custom	1	1	46
	USR		80	0.813	
	SHAFTS		not found	–	
wwLig-CSRre	LigCSR	DrugBank Approved	not found	–	47
iDrug	SHAFTS	MayBridge	not found	–	48
		ZINC Lead	not found	–	

<sup>a</sup>The full list of experiments performed on BRUSELAS is provided in the Supporting Information.



Figure 7. Alignments obtained for acarbose in the simulation combining the four algorithms.

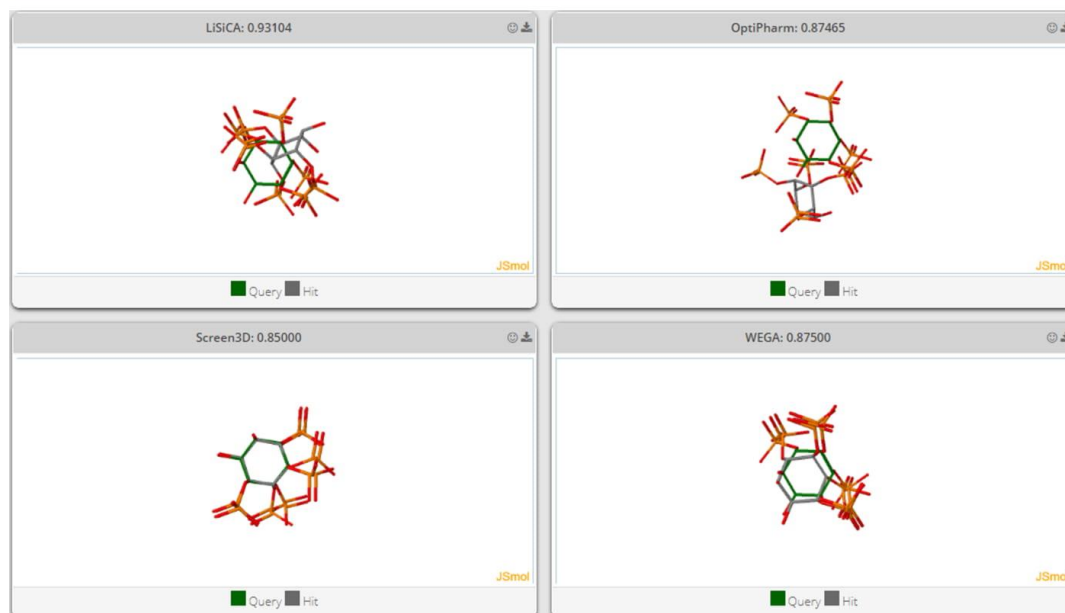
to 1, which represents total similarity in terms of the Tanimoto coefficient.

We split the tests into two groups: shape similarity and pharmacophore searching tasks. As regards shape similarity, Table 2 (see full details in Supporting Information) shows that BRUSELAS correctly predicted acarbose as the best-scored compound in all the searches, independent of the configuration chosen, resolving each task in less than half an hour. Figure 7 shows the alignments obtained for acarbose from all the

algorithms. As regards of the other servers, they can be classified into two groups, depending on whether they identified acarbose in the candidates list or not. The first group includes SwissSimilarity, HybridSim-VS, and ChemMapper. Although SwissSimilarity and ChemMapper exhibited the highest scores, the former did not find any result when screening with Shape-IT and Align-IT, and the latter only performs 2D screening and did not rank acarbose as the first option when using USR. On the other hand, HybridSim-VS

**Table 3.** Summary of the most relevant results obtained by BRUSELAS for TMI query using similarity and pharmacophore searches

similarity method	compound	score	ref
shape similarity	phosphoric acid mono((1 <i>S</i> ,2 <i>S</i> ,3 <i>R</i> ,4 <i>S</i> ,5 <i>R</i> ,6 <i>R</i> )-2,3-dihydroxy-4,5,6-tris(phosphonoxy)cyclohexyl) ester	0.883	57
	phosphoric acid mono((1 <i>R</i> ,2 <i>R</i> ,3 <i>S</i> ,4 <i>R</i> ,5 <i>S</i> ,6 <i>S</i> )-2,3-dihydroxy-4,5,6-tris(phosphonoxy)cyclohexyl) ester	0.811	
	ionositol 1,3,4,6-tetraphosphate	0.771	
	phosphoric acid mono((1 <i>S</i> ,2 <i>S</i> ,3 <i>R</i> ,4 <i>S</i> ,5 <i>S</i> ,6 <i>S</i> )-2,5-dihydroxy-3,4,6-tris(phosphonoxy)cyclohexyl) ester	0.751	
Pharmacophore screening	L-adenosine triphosphate	0.567	58
	D-myo-inositol 1,2,4,5-tetrakisphosphate	0.233	
	2-amino-9-(4-hydroxy-5-methoxytriphosphate tetrahydrofuran-2-yl)-1,9-dihydropurin-6-one (dGTP)	0.185	
	{{[( <i>S</i> )-{[( <i>S</i> )-{[(2 <i>R</i> ,3 <i>S</i> ,4 <i>R</i> ,5 <i>R</i> )-5-(6-amino-9 <i>H</i> -purin-9-yl)-3,4-dihydroxyoxolan-2-yl]methoxy}(hydroxy)phosphoryl]amino}phosphonic acid oxy}(hydroxy)phosphoryl]amino}phosphonic acid	0.183	
	((2 <i>R</i> ,3 <i>S</i> ,4 <i>R</i> ,5 <i>R</i> )-5-(7-amino-3 <i>H</i> -[1,2,3]triazolo[4,5- <i>d</i> ]pyrimidin-3-yl)-3,4-dihydroxytetrahydrofuran-2-yl)methyltriphosphoric acid	0.180	
	ATP analog	0.167	

**Figure 8.** Alignment of TMI and phosphoric acid mono((1*S*,2*S*,3*R*,4*S*,5*R*,6*R*)-2,3-dihydroxy-4,5,6-tris(phosphonoxy)cyclohexyl) ester obtained from the different algorithms.

always ranked acarbose as the best result, but the scores were far from those expected (0.693 was the highest score). We can see that BRUSELAS succeeded in all of the cases, and it gave acarbose higher scores than HybridSim-VS due to use of a consensus scoring function. The second group contains USR-VS, Superimposé and wwLig-CSRre, which either did not return any result or did not even finish the calculations with some of the configurations. The behavior of Superimposé may be explained by the fact that it only screens its own databases, which probably do not contain acarbose. USR-VS performed best in terms of computation time, but even though it does not screen the FDA approved drugs data set, it was expected to find acarbose in ZINC database because the query molecule was extracted from that database. Finally, wwLig-CSRre did not find the expected molecule when screening the same DrugBank subset as HybridSim-VS, suggesting a difference in the similarity algorithm used by both servers.

In terms of pharmacophore screening, all the servers providing this feature use the SHAFTS algorithm to assess similarity. It is observed that none of the servers (iDrug, ChemMapper and BRUSELAS) found acarbose in the list of hits, which suggests that the algorithm is unable to find the match.

**4.3. Blood Anticoagulants.** The medical challenge faced by an aging society with its need for anticoagulant drugs of increasing complexity encourages the search for new anticoagulant molecules. Heparin is widely used as an activator of antithrombin, although it has side effects. In a previous virtual screening campaign we discovered<sup>55,56</sup> that the novel compound D-myo-inositol 3,4,5,6-tetrakisphosphate (TMI) is able to act as a heparin cofactor, binding with nanomolar affinity to antithrombin and causing its partial activation. Given its novel scaffold, it paves the way for the discovery of novel anticoagulants based on its molecular structure. Using

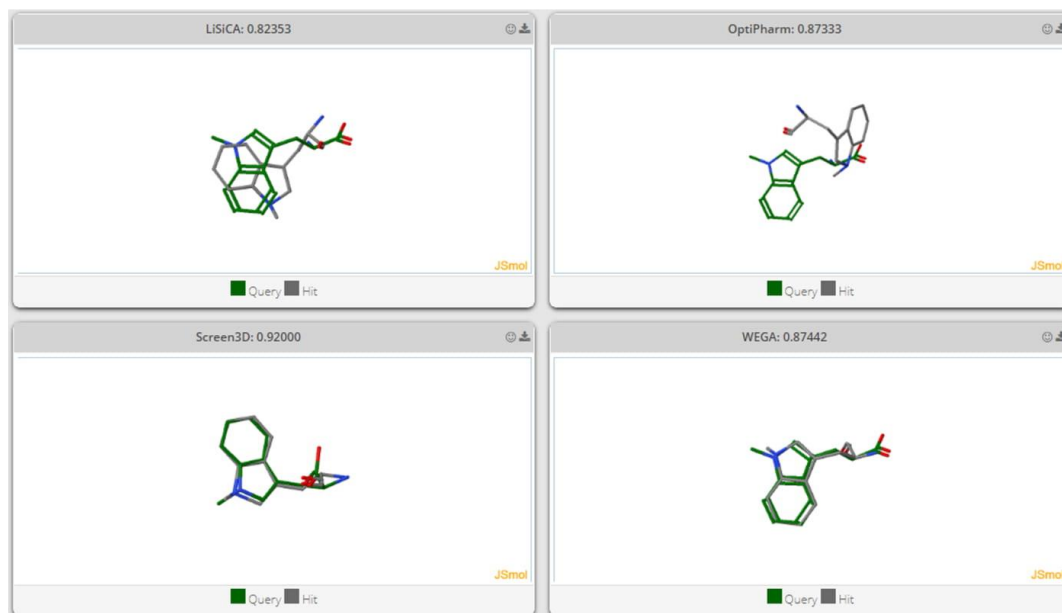


Figure 9. Alignment of Indoximod and (S)-2-amino-3-(1-methyl-1H-indol-3-yl)propanal obtained from the different algorithms.

BRUSELAS, we tested TMI using it as a query structure and performing both shape and pharmacophore searches against the ChEMBL database. Table 3 shows the most relevant hits obtained among the top scoring compounds. The full list of results is provided in Supporting Information.

Among the hits obtained by both methods were many different isomers of TMI, as was to be expected. For example, Figure 8 shows the alignment of TMI against one of the TMI isomers mentioned. It might be practical, then, to test these compounds experimentally as potential anticoagulants. Also of note was the fact that the method found ADP, which has previously been reported to act as a blood anticoagulant.<sup>59</sup> Other ADP related compounds, such as dGTP which has not previously been reported, might also be worth testing.

**4.4. IDO1.** The above examples looked at some representative uses of computational tools with a focus on drug discovery. Such approaches may help to rationalize the synthesis of new drugs for cancer therapy.<sup>60</sup> Indeed, there are several successful cases that target different steps in the cascade of reactions involved in tumor propagation, e.g., the tumor suppressor protein p53, proteins of the signal transducer and activator of transcription or molecular switches that control signaling pathways, to cite a few.<sup>61,62</sup>

In spite of such efforts, patients with metastatic cancer represent one of the most urgent and challenging goals for medicine, where available therapies fail because malignant cells become drug-resistant during treatment. One of the most relevant achievements in this line has been the discovery of the cytosolic enzyme indoleamine 2,3-dioxygenase-1 (IDO1).<sup>63</sup> According to recent preclinical results, the inhibition of the IDO1 enzyme enhances the efficacy of classical chemotherapy, radiotherapy, and immune checkpoint therapy while bypassing their side effects.<sup>64</sup> This encouraging evidence, in turn, has led to the search for novel inhibitors by using both experimental

and theoretical methods.<sup>65,66</sup> For the latter, Zheng, Yan and co-workers have recently used SBVS methods to propose novel IDO1 inhibitors.<sup>65</sup> However, to the best of our knowledge, there has been no systematic search using LBVS, so that this biological problem can be regarded as an optimal working example to test our software architecture. We selected three queries from the IDO1 inhibitors reported by the pharmaceutical industry, namely, indoximod, navoximod, and epacadostat.<sup>67</sup>

It is of note that (S)-2-amino-3-(1-methyl-1H-indol-3-yl)propanal, whose alignments with indoximod are displayed in Figure 9, is located at the top of the list, as this molecule is known to be a potent inhibitory activity on IDO1, as demonstrated by Frédéric and co-workers, who combined docking predictions and *in vivo* experiments.<sup>68</sup> Our calculations, which were performed without imposing any bias or restriction, were led to similar results with a significantly lighter computational effort. Less expected were the rest of the compounds arising from both shape similarity and pharmacophore screening. For instance, (S)-methyl 2-amino-3-(1H-indol-3-yl)propanoate hydrochloride has been shown to act on Gap1, but to the best of our knowledge, it has not previously been tested in that context. We therefore conclude that the presented server correctly identified a known IDO1 inhibitor, which benchmarks the implemented protocol. Simultaneously, it is our hope that the new compound included in Table 4 may be used to expand the chemical space of novel and more efficient inhibitors beyond the more classical indoximod, navoximod, and epacadostat (see further details and results in the Supporting Information).

## 5. CONCLUSIONS AND FUTURE WORK

This contribution presents a novel software architecture, BRUSELAS, which performs 3D LBVS based on shape

**Table 4. Best-Ranked IDO1 Inhibitors Proposed by BRUSELAS Using Indoximod, Navoximod, and Epacadostat as Queries on Shape and Pharmacophore Searches**

similarity method	compound	score
shape similarity	(S)-2-amino-3-(1-methyl-1H-indol-3-yl)propanal	0.873
	(S)-methyl 2-amino-3-(1H-indol-3-yl)propanoate hydrochloride	0.819
	3-amino-2-(1H-indol-3-yl)-propionic acid methyl ester hydrochloride	0.809
	CHEMBL162163	0.809
	3-(2-(1H-tetrazol-5-yl)ethyl)-1H-indole	0.772
pharmacophore screening	2-amino-3-(1H-indol-3-yl)propionic acid methyl ester	0.308
	(S)-methyl 2-amino-3-(1H-indol-3-yl)propanoate hydrochloride	0.307
	SID103076391	0.302
	1-benzyl-2-propyl-1H-imidazo[4,5-c]quinolin-4-amine	0.266
	5,N*6*-Dimethyl-N*6*-(2-methylpyridin-4-ylmethyl)benzo[cd]indole-2,6-diamine	0.263

similarity or pharmacophore searching. It also maintains a large collection of compounds to suggest a fitting library of ligands for the given query molecule. Aiming to satisfy researchers' needs, BRUSELAS intends to be a powerful platform for the application of LBVS techniques to drug discovery. To achieve this aim, the architecture exhibits a robust and flexible design complemented with a web tool to facilitate setting tasks and analyzing the results. BRUSELAS possesses distinguishing features that make it innovative in the field of LBVS. Those features are represented by the combined use of many similarity algorithms in the same task, the application of consensus scoring functions and the selection of keywords to focus searches on the desired disease or target. Additionally, tests proved that BRUSELAS outperforms many other similar servers in terms of reliability. It was one of the two servers succeeding in all the shape similarity searches, and it reached a good balance between speed and accuracy in comparison with others. Moreover, its applicability to different contexts has been demonstrated by searching blood anticoagulants and IDO1 enzyme inhibitors.

The comparison with similar servers suggests that BRUSELAS is of potential application in many other virtual screening studies. Nevertheless, some remaining challenges might be worth further study in the future. A more precise algorithm for choosing the most similar compounds based on the selected descriptors and the automatic assignment of weights to similarity algorithms are examples of such challenges. Moreover, a revision of the existing and newly published algorithms and an update of the libraries, which is a task that takes several months, have to be carried out regularly. Despite this, BRUSELAS has been demonstrated to be effective for the identification of new active molecules with innovative chemical scaffolds and to be reliable enough for experimental validation.

## ■ ASSOCIATED CONTENT

### ● Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00279.

Table S1, comparison of a collection of similarity web servers performing similarity searches with acarbose as the query molecule, which is an extension of Table 2,

containing all the experiments performed with BRUSELAS, considering the different similarity algorithms; Table S2, full list of candidates returned by the server when providing TMI as the query molecule, which is an extended version of Table 3; Table S3, complete list of candidates returned by BRUSELAS while looking for IDO1 inhibitors, contains similarities with indoximod, navoximod, and epacadostat and extends Table 4 (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

\*(A.J.B.-L.) E-mail: [ajbanegas@alu.ucam.edu](mailto:ajbanegas@alu.ucam.edu).

\*(H.P.-S.) E-mail: [hperez@ucam.edu](mailto:hperez@ucam.edu).

### ORCID

Antonio J. Banegas-Luna: 0000-0003-1158-8877

José P. Cerón-Carrasco: 0000-0003-0668-9227

Savins Puertas-Martín: 0000-0001-8956-1733

Horacio Pérez-Sánchez: 0000-0003-4468-7898

### Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

### Funding

This work was funded by grants from the Spanish Ministry of Economy and Competitiveness (CTQ2017-87974-R and RTI2018-095993-B-I00) and by the Fundación Séneca del Centro de Coordinación de la Investigación de la Región de Murcia under Projects 20988/PI/18 and 20524/PDC/18. This research was partially supported by the supercomputing infrastructure of Poznan Supercomputing Center, the e-infrastructure program of the Research Council of Norway, the supercomputer center of UiT – the Arctic University of Norway and by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain. The authors also acknowledge the computing resources and technical support provided by the Plataforma Andaluza de Bioinformática de the University of Málaga. Powered@NLHPC research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02). Savins Puertas Martín is a fellow of the Spanish 'Formación del Profesorado Universitario' program, financed by the Spanish Ministry of Education, Culture and Sport.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was funded by grants from the Spanish Ministry of Economy and Competitiveness (CTQ2017-87974-R and RTI2018-095993-B-I00) and by the Fundación Séneca del Centro de Coordinación de la Investigación de la Región de Murcia under Projects 20988/PI/18 and 20524/PDC/18. This research was partially supported by the supercomputing infrastructure of Poznan Supercomputing Center, the e-infrastructure program of the Research Council of Norway, the supercomputer center of UiT – the Arctic University of Norway and by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the

Government of Spain. The authors also acknowledge the computing resources and technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga. Powered@NLHPC research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02). Savins Puertas Martín is a fellow of the Spanish 'Formación del Profesorado Universitario' program, financed by the Spanish Ministry of Education, Culture and Sport.

#### ■ ABBREVIATIONS

ADP, adenosine diphosphate; BRUSELAS, balanced rapid and unrestricted server for extensive ligand-aimed screening; FDA, Food and Drug Administration; HPC, high-performance computing; IDO1, indoleamine 2,3-dioxygenase-1; IUPAC, International Union of Pure and Applied Chemistry; LBVS, ligand-based virtual screening; MMFF94, Merck molecular force field; QSAR, quantitative structure–activity relationship; RMSD, root-mean-squared distance; SBVS, structure-based virtual screening; SMILES, simplified molecular input line entry specification; T2DM, Type 2 Diabetes Mellitus; TMI, D-myo-inositol 3,4,5,6-tetrakisphosphate; VS, virtual screening.

#### ■ REFERENCES

(1) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303* (5665), 1813–1818.

(2) Dou, X.; Jiang, L.; Wang, Y.; Jin, H.; Liu, Z.; Zhang, L. Discovery of new GSK-3 $\beta$  inhibitors through structure-based virtual screening. *Bioorg. Med. Chem. Lett.* **2018**, *28* (2), 160–166.

(3) Nunes, R. R.; dos Santos Costa, M.; dos Reis Santos, B.; da Fonseca, A. L.; Ferreira, L. S.; Russo Chagas, R. C.; da Silva, A. M.; de Pilla Varotti, F.; Taranto, A. G. Successful application of virtual screening and molecular dynamics simulations against antimalarial molecular targets. *Mem. Inst. Oswaldo Cruz* **2016**, *111* (12), 721–730.

(4) Tian, S. T.; Wang, X.; Li, L.; Zhang, X.; Li, Y.; Zhu, F.; Hou, T.; Zhen, X. Discovery of novel and selective adenosine A2A receptor antagonists for treating Parkinson's disease through comparative structure-based virtual screening. *J. Chem. Inf. Model.* **2017**, *57* (6), 1474–1487.

(5) Tikhonova, I. G.; Sum, C. S.; Neumann, S.; Engel, S.; Raaka, B. M.; Costanzi, S.; Gershengorn, M. C. Discovery of novel agonists and antagonists of the free fatty acid receptor one (FFAR1) using virtual screening. *J. Med. Chem.* **2008**, *51* (3), 625–633.

(6) Schuster, D.; Maurer, E. M.; Laggner, C.; Nashev, L. G.; Wilckens, T.; Langer, T.; Odermatt, A. The discovery of new 11 $\beta$ -hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J. Med. Chem.* **2006**, *49* (12), 3454–3466.

(7) Salam, N. K.; Huang, T. H.; Kota, B. P.; Kim, M. S.; Li, Y.; Hibbs, D. E. Novel PPAR- $\gamma$  agonists identified from a natural product library: a virtual screening, induced-fit docking and biological assay study. *Chem. Biol. Drug Des.* **2008**, *71* (1), 57–70.

(8) Mahmoudi, N.; de Julián-Ortiz, J. V.; Ciceron, L.; Gálvez, J.; Mazier, D.; Danis, M.; Derouin, F.; García-Domenech, R. Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening. *J. Antimicrob. Chemother.* **2006**, *57* (3), 489–497.

(9) Desai, P. V.; Patny, A.; Sabnis, Y.; Tekwani, B.; Gut, J.; Rosenthal, P.; Srivastava, A.; Avery, M. Identification of novel parasitic cysteine protease inhibitors using virtual screening. 1. The ChemBridge database. *J. Med. Chem.* **2004**, *47* (26), 6609–6615.

(10) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48* (7), 2534–2547.

(11) Evers, A.; Klabunde, T. Structure-based Drug Discovery Using GPCR Homology Modeling: Successful Virtual Screening for Antagonists of the Alpha1A Adrenergic Receptor. *J. Med. Chem.* **2005**, *48* (4), 1088–1097.

(12) Rollinger, J. M.; Hornick, A.; Langer, T.; Stuppner, H.; Prast, H. Acetylcholinesterase Inhibitory Activity of Scopolin and Scopoletin Discovered by Virtual Screening of Natural Products. *J. Med. Chem.* **2004**, *47* (25), 6248–6254.

(13) Vogt, M.; Bajorath, J. Predicting the performance of fingerprint similarity searching. *Methods Mol. Biol.* **2010**, *672*, 159–173.

(14) Kumar, A.; Zhang, K. Y. J. Hierarchical virtual screening approaches in small molecule drug discovery. *Methods* **2015**, *71*, 26–37.

(15) Kaserer, T.; Beck, K. R.; Akram, M.; Odermatt, A.; Schuster, D. Pharmacophore models and pharmacophore-based virtual screening: concepts and applications exemplified on hydroxysteroid dehydrogenases. *Molecules* **2015**, *20*, 22799–22832.

(16) Villoutreix, B. O.; Lagorce, D.; Labbé, C. M.; Sperandio, O.; Miteva, M. A. One hundred thousand mouse clicks down the road: Selected online resources supporting drug discovery collected over a decade. *Drug Discovery Today* **2013**, *18* (21–22), 1081–1089.

(17) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing pitfalls in virtual screening: a critical review. *J. Chem. Inf. Model.* **2012**, *52* (4), 867–881.

(18) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2009.

(19) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for *in Silico* Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34* (90001), D668–D672.

(20) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100–D1107.

(21) Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44* (D1), D457–D462.

(22) DIA-DB; Bioinformatics and High Performance Computing Research Group: 2015.

(23) ChemAxon; ChemAxon Ltd., C.

(24) Omega, OpenEye; OpenEye: 2016.

(25) Yan, X.; Li, J.; Liu, Z.; Zheng, M.; Ge, H.; Xu, J. Enhancing Molecular Shape Comparison by Weighted Gaussian Functions. *J. Chem. Inf. Model.* **2013**, *53* (8), 1967–1978.

(26) Lesnik, S.; Stular, T.; Brus, B.; Knez, D.; Gobec, S.; Janezic, D.; Konc, J. LiSiCA: A Software for Ligand-Based Virtual Screening and Its Application for the Discovery of Butyrylcholinesterase Inhibitors. *J. Chem. Inf. Model.* **2015**, *55* (8), 1521–1528.

(27) Kalaszi, A.; Szisz, D.; Imre, G.; Polgar, T. Screen3D: A Novel Fully Flexible High-Throughput Shape-Similarity Search Method. *J. Chem. Inf. Model.* **2014**, *54* (4), 1036–1049.

(28) Puertas-Martin, S.; Redondo, J. L.; Pérez-Sánchez, H.; Ortigosa, P. M. OptiPharm: An evolutionary algorithm to compare shape similarity. *Sci. Rep.* **2019**, *9*, 1398.

(29) Liu, X.; Jiang, H.; Li, H. SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. *J. Chem. Inf. Model.* **2011**, *51* (9), 2372–2385.

(30) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3* (1), 33.

(31) Hanson, R. M. Jmol - a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr.* **2010**, *43* (5), 1250–1260.

(32) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5–6), 490–519.

(33) Lu, W.; Liu, X.; Cao, X.; Xue, M.; Liu, K.; Zhao, Z.; Shen, X.; Jiang, H.; Xu, Y.; Huang, J.; Li, H. SHAFTS: A hybrid approach for 3D molecular similarity calculation. 2. Prospective case study in the



- discovery of diverse p90 ribosomal S6 protein kinase 2 inhibitors to suppress cell migration. *J. Med. Chem.* **2011**, *54* (10), 3564–3574.
- (34) Kong, X.; Qin, J.; Li, Z.; Vultur, A.; Tong, L.; Feng, E.; Rajan, G.; Liu, S.; Lu, J.; Liang, Z.; Zheng, M.; Zhu, W.; Jiang, H.; Herlyn, M.; Liu, H.; Marmorstein, R.; Luo, C. Development of a novel class of B-Raf(V600E)-selective inhibitors through virtual screening and hierarchical hit optimization. *Org. Biomol. Chem.* **2012**, *10* (36), 7402–7417.
- (35) Leelananda, S. P.; Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **2016**, *12*, 2694–2718.
- (36) Jaghoori, M. M.; Bleijlevens, B.; Olabbariaga, S. D. 1001 Ways to Run AutoDock Vina for Virtual Screening. *J. Comput.-Aided Mol. Des.* **2016**, *30* (3), 237–249.
- (37) Please, ensure you add biohpc2015@gmail.com to your whitelist. If you do not receive a notification mail within 48 h after the creation of the job, check your spam folder.
- (38) Esu, E.; Effa, E. E.; Opie, O. N.; Uwaoma, A.; Meremikwu, M. M. Artemether for Severe Malaria (Review). *Cochrane Libr* **2014**, No. 9, 1–85.
- (39) Teklemariam, M.; Assefa, A.; Kassa, M.; Mohammed, H.; Mamo, H. Therapeutic Efficacy of Artemether-Lumefantrine against Uncomplicated Plasmodium Falciparum Malaria in a High-Transmission Area in Northwest Ethiopia. *PLoS One* **2017**, *12* (4), e0176004.
- (40) Phyto, A. P.; von Seidlein, L. Challenges to Replace ACT as First-line Drug. *Malar. J.* **2017**, *16* (1), 296.
- (41) Plucinski, M. M.; Ferreira, M.; Ferreira, C. M.; Burns, J.; Gaparayi, P.; João, L.; da Costa, O.; Gill, P.; Samutondo, C.; Quivinja, J.; Mbounga, E.; de León, G. P.; Halsey, E. S.; Dimbu, P. R.; Fortes, F. Evaluating Malaria Case Management at Public Health Facilities in Two Provinces in Angola. *Malar. J.* **2017**, *16* (1), 186.
- (42) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. USR-VS: A Web Server for Large-Scale Prospective Virtual Screening Using Ultrafast Shape Recognition Techniques. *Nucleic Acids Res.* **2016**, *44* (W1), W436–441.
- (43) Zoete, V.; Daina, A.; Bovigny, C.; Michielin, O. SwissSimilarity: A Web Tool for Low to Ultra High Throughput Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2016**, *56* (8), 1399–1404.
- (44) Bauer, R. A.; Bourne, P. E.; Formella, A.; Frömmel, C.; Gille, C.; Goede, A.; Guerler, A.; Hoppe, A.; Knapp, E. W.; Pöschel, T.; Wittig, B.; Ziegler, V.; Preissner, R. Superimposé: A 3D Structural Superposition Server. *Nucleic Acids Res.* **2008**, *36* (WebServer), W47–W54.
- (45) Shang, J.; Dai, X.; Li, Y.; Pistozzi, M.; Wang, L. HybridSimVS: a web server for large-scale ligand-based virtual screening using hybrid similarity recognition techniques. *Bioinformatics* **2017**, *33* (21), 3480–3481.
- (46) Gong, J.; Cai, C.; Liu, X.; Ku, X.; Jiang, H.; Gao, D.; Li, H. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics* **2013**, *29* (14), 1827–1829.
- (47) Sperandio, O.; Petitjean, M.; Tuffery, P. wwLigCSRre: A 3D ligand-based server for hit identification and optimization. *Nucleic Acids Res.* **2009**, *37* (WebServer), W504–W509.
- (48) Wang, X.; Chen, H.; Yang, F.; Gong, J.; Li, S.; Pei, J.; Liu, X.; Jiang, H.; Lai, L.; Li, H. iDrug: a web-accessible and interactive drug discovery and design platform. *J. Cheminf.* **2014**, *6* (1), 28.
- (49) Ge, Q.; Chen, L.; Chen, K. Treatment of Diabetes Mellitus Using iPSC Cells and Spice Polyphenols. *J. Diabetes Res.* **2017**, *2017*, 1–11.
- (50) Liu, Z.; Zhao, X.; Sun, W.; Wang, Y.; Liu, S.; Kang, L. E. I. Metformin Combined with Acarbose vs. Single Medicine in the Treatment of Type 2 Diabetes: A Meta-Analysis. *Exp. Ther. Med.* **2017**, *13*, 3137–3145.
- (51) Wettergreen, S. A.; Sheth, S.; Malveaux, J. Effects of the Addition of Acarbose to Insulin and Non-Insulin Regimens in Veterans with Type 2 Diabetes Mellitus. *Pharm. Pract. (Granada)*. **2016**, *14* (4), 832.
- (52) Cadegiani, F. A.; Silva, O. S. Acarbose Promotes Remission of Both Early and Late Dumping Syndromes in Post-Bariatric Patients. *Diabetes, Metab. Syndr. Obes.: Targets Ther.* **2016**, *9*, 443–446.
- (53) Li, J.; Niu, B.; Wang, X.; Hu, H.; Cao, B. A Case Report of Hereditary Neuropathy with Liability to Pressure Palsies Accompanied by Type 2 Diabetes Mellitus and Psoriasis. *Medicine* **2017**, *96* (19), e6922.
- (54) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52* (7), 1757–1768.
- (55) Adelt, S.; Plettenburg, O.; Stricker, R.; Reiser, G.; Altenbach, H. J.; Vogel, G. Enzyme-Assisted Total Synthesis of the Optical Antipodes d-myo-Inositol 3,4,5-Trisphosphate and d-myo-Inositol 1,5,6-Trisphosphate: Aspects of Their Structure–Activity Relationship to Biologically Active Inositol Phosphates. *J. Med. Chem.* **1999**, *42* (7), 1262–1273.
- (56) Navarro-Fernández, J.; Pérez-Sánchez, H.; Martínez-Martínez, I.; Melicani, I.; Guerrero, J. A.; Vicente, V.; Corral, J.; Wenzel, W. In Silico Discovery of a Compound with Nanomolar Affinity to Antithrombin Causing Partial Activation and Increased Heparin Affinity. *J. Med. Chem.* **2012**, *55* (14), 6403–6412.
- (57) Hocek, M.; Silhár, P.; Shih, I. H.; Mabery, E.; Mackman, R. Cytostatic and antiviral 6-arylpurine ribonucleosides. Part 7: Synthesis and evaluation of 6-substituted purine l-ribonucleosides. *Bioorg. Med. Chem. Lett.* **2006**, *16* (20), 5290–5293.
- (58) Quick, A. J. Anticoagulant Action of Adenosine Triphosphate. *Nature* **1963**, *200*, 469–470.
- (59) Wassman, C. D.; Baronio, R.; Demir, Ö.; Wallentine, B. D.; Chen, C. K.; Hall, L. V.; Salehi, F.; Lin, D. W.; Chung, B. P.; Hatfield, G. W.; Richard Chamberlin, A.; Luecke, H.; Lathrop, R. H.; Kaiser, P.; Amaro, R. E. Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nat. Commun.* **2013**, *4* (1), 1407.
- (60) Siddiquee, K.; Zhang, S.; Guida, W. C.; Blaskovich, M. A.; Greedy, B.; Lawrence, H. R.; Yip, M. L.; Jove, R.; McLaughlin, M. M.; Lawrence, N. J.; Sebt, S. M.; Turkson, J. Selective chemical probe inhibitor of Stat3, identified through structure-based virtual screening, induces antitumor activity. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (18), 7391–7396.
- (61) Gao, Y.; Dickerson, J. B.; Guo, F.; Zheng, J.; Zheng, Y. Rational design and characterization of a Rac GTPase-specific small molecule inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (20), 7618–7623.
- (62) Muller, A. J.; DuHadaway, J. B.; Donover, P. S.; Sutanto-Ward, E.; Prendergast, G. C. Inhibition of indoleamine 2,3-dioxygenase, an immunoregulatory target of the cancer suppression gene Bin1, potentiates cancer chemotherapy. *Nat. Med.* **2005**, *11*, 312–319.
- (63) Hou, D. Y.; Muller, A. J.; Sharma, M. D.; DuHadaway, J.; Banerjee, T.; Johnson, M.; Mellor, A. L.; Prendergast, G. C.; Munn, D. H. Inhibition of indoleamine 2,3-dioxygenase in dendritic cells by stereoisomers of 1-methyl-tryptophan correlates with antitumor responses. *Cancer Res.* **2007**, *67* (2), 792–801.
- (64) Godin-Ethier, J.; Hanafi, L. A.; Piccirillo, C. A.; Lapointe, R. Indoleamine 2,3-dioxygenase expression in human cancers: clinical and immunologic perspectives. *Clin. Cancer Res.* **2011**, *17* (22), 6985–6991.
- (65) Zhang, G.; Xing, J.; Wang, Y.; Wang, L.; Ye, Y.; Lu, D.; Zhao, J.; Luo, X.; Zheng, M.; Yan, S. Discovery of Novel Inhibitors of Indoleamine 2,3-Dioxygenase 1 Through Structure-Based Virtual Screening. *Front. Pharmacol.* **2018**, *9*, 00277.
- (66) Prendergast, G. C.; Malachowski, W. P.; DuHadaway, J. B.; Muller, A. J. Discovery of IDO1 Inhibitors: From Bench to Bedside. *Cancer Res.* **2017**, *77* (24), 6795–6811.
- (67) Navarro-Fernández, J.; Pérez-Sánchez, H.; Martínez-Martínez, I.; Melicani, I.; Guerrero, J. A.; Vicente, V.; Corral, J.; Wenzel, W. In Silico Discovery of a Compound with Nanomolar Affinity to Antithrombin Causing Partial Activation and Increased Heparin Affinity. *J. Med. Chem.* **2012**, *55* (14), 6403–6412.
- (68) Dolušić, E.; Larrieu, P.; Moineaux, L.; Stroobant, V.; Pilotte, L.; Colau, D.; Pochet, L.; Van den Eynde, B.; Masereel, B.; Wouters, J.

Frédéric, R. Tryptophan 2,3-Dioxygenase (TDO) Inhibitors. 3-(2-(Pyridyl)ethenyl)indoles as Potential Anticancer Immunomodulators. *J. Med. Chem.* **2011**, *54* (15), 5320–5334.

## **IV - CONCLUSIONES**



## IV - CONCLUSIONES

Los artículos que componen el compendio de esta tesis examinan la necesidad de desarrollar una herramienta como BRUSELAS, a la vez que desglosan sus principales características y modo de empleo. Esta sección expone las principales conclusiones obtenidas a lo largo de este trabajo, y propone una serie líneas de investigación para el futuro que deben marcar el camino a seguir para la evolución y optimización de la nueva arquitectura.

### 4.1 CONCLUSIONES

Del análisis de las herramientas LBVS se desprenden algunas conclusiones acerca de las funcionalidades que ofrecen a los usuarios.

1. Se puede observar que los algoritmos de similitud han evolucionado desde complejos programas de línea de comandos hasta ser empleados a través de servidores web, con el objetivo de facilitar y acercar su uso a usuarios menos expertos.
2. Además, a medida que el big data se ha hecho un hueco en la ciencia actual, los servidores de LBVS se han visto en la necesidad de procesar volúmenes de datos más y más grandes para lo que han optado por limitar el número y el tamaño de las librerías disponibles para VS, y al empleo de plataformas HPC para poder reunir la potencia de cómputo necesaria.
3. Esta evolución en los algoritmos y en su modo de distribución hace pensar que una arquitectura para LBVS debe ser de aplicabilidad general, accesible a través de un entorno web y estar apoyada en alguna infraestructura HPC para conseguir un rendimiento competitivo.

BRUSELAS no sólo reúne todas las características anteriormente mencionadas, sino que incluye funcionalidades adicionales que le hacen aportar un valor científico adicional. Dichas características son el consenso de algoritmos de similitud, la creación de librerías dinámicas a partir de un conjunto de

descriptores, la utilización de palabras clave para seleccionar los compuestos a incluir en las librerías y el empleo de filtros moleculares para excluir los resultados indeseados. Todas las propiedades mencionadas han sido utilizadas de manera retrospectiva en la búsqueda de anticoagulantes sanguíneos, antidiabéticos e inhibidores de la enzima IDO1 que participa en diversas terapias para el tratamiento del cáncer. Los resultados obtenidos, aun siendo teóricos, son prometedores, tanto en búsquedas por similitud como farmacofóricas.

4. Estos resultados prueban que las nuevas funcionalidades de BRUSELAS son efectivas para realizar un cribado eficaz en las etapas iniciales de la búsqueda de compuestos líder.
5. Se puede concluir, en consecuencia, que librerías las químicas adaptadas a cada molécula de referencia permiten cribar pequeñas librerías muy concretas para cada problema específico, pero que son creadas dinámicamente. Además, el consenso de las puntuaciones corrige las posibles desviaciones causadas por cada algoritmo individual lo que incrementa la fiabilidad de las predicciones de manera notable.

En cuanto al rendimiento en términos computacionales se puede concluir que:

6. Nuestro servidor está al mismo nivel que sus competidores, incluso cuando debe cribar más de 7,5 millones de compuestos para crear una librería más reducida, evaluar la similitud 3D entre la molécula de referencia y cada ligando, y aplicar la función de consenso para calcular la puntuación final. Los únicos casos en los que BRUSELAS no obtuvo resultados fueron debidos a las limitaciones de los algoritmos subyacentes como se explica en el tercer artículo del compendio.
7. En consecuencia, se puede afirmar que BRUSELAS es una alternativa eficaz y eficiente para tareas de cribado virtual basado en ligandos debido a su buen rendimiento, a que evita desviaciones causadas por el uso un único algoritmo y a que es capaz de cribar librerías específicamente creadas para cada molécula de entrada.

## 4.2 FUTURAS LÍNEAS DE INVESTIGACIÓN

La existencia de un espacio químico cambiante y la evolución de los algoritmos de VS hacen que se haga necesaria una revisión frecuente de la información que maneja BRUSELAS para mantener sus predicciones actualizadas. En esta sección se exponen éstas y otras posibles líneas de investigación futuras sobre la herramienta desarrollada.

Es de esperar que, en el futuro, aparezcan nuevos y mejores algoritmos de similitud por forma y farmacofórica, realizando mejores predicciones. Con la intención de que BRUSELAS sea una herramienta flexible que se adapte a la tecnología del momento, sería muy interesante ir completando el conjunto de algoritmos de similitud con los que aparezcan en el futuro. Esta tarea debe incluir tanto la adición de nuevos algoritmos, como la actualización de los actuales a versiones más optimizadas. De igual manera, es de prever que las bases de datos importadas también evolucionen a gran velocidad, lo que requerirá el mantenimiento y la actualización de la base de datos de compuestos de BRUSELAS. Paralelamente, habrá que actualizar la lista de términos clave asociados a cada uno de ellos, con el fin de crear librerías actualizadas en todo momento.

BRUSELAS cubre dos de los tres tipos de LBVS más importantes: la búsqueda por similitud y el modelado farmacofórico. Un punto muy interesante a explorar sería completar esta lista de técnicas con métodos QSAR. Siguiendo la línea de trabajo de la arquitectura, se crearía una librería adaptada a cada caso y se aplicaría una función de consenso sobre los resultados. Es de esperar que, al igual que ocurre en las otras técnicas, el consenso de puntuaciones dé lugar a predicciones sin sesgo y más realistas.

Una tercera línea de investigación sería la ejecución de una campaña de VS completa en la que, primero, se obtiene un conjunto de líderes candidatos mediante BRUSELAS y, posteriormente, esos compuestos son utilizados como entrada de una herramienta de *docking*. Este experimento tendría un doble valor: corroborar que la colección de líderes devuelta por nuestra arquitectura es verdaderamente válida, y obtener información más detallada de cómo esos ligandos se acoplan al receptor estudiado.

Todas estas líneas de investigación están encaminadas a mantener la herramienta siempre actualizada a la información disponible en cada momento, y a completar sus predicciones con otras técnicas más costosas computacionalmente, las cuales sólo son eficientes si trabajan sobre pequeños volúmenes de datos.



## **V – REFERENCIAS BIBLIOGRÁFICAS**



**V – REFERENCIAS BIBLIOGRÁFICAS**

1. Murray CJ, Ortblad KF, Guinovart C, Lim SS, Wolock TM, Roberts DA, et al. Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2014; 384(9947):1005-70.
2. Vos T, Barber RM, Bell B, Bertozzi-Villa A, Biryukov S, Bolliger I, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015; 386(9995):743-800.
3. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015; 136(5):E359-86.
4. Leszek J, Md Ashraf G, Tse WH, Zhang J, Gasiorowski K, Avila-Rodriguez MF, et al. Nanotechnology for Alzheimer Disease. *Curr Alzheimer Res*. 2017; 14(11):1182-89.
5. Abbruzzese G, Marchese R, Avanzino L, Pelosin E. Rehabilitation for Parkinson's disease: Current outlook and future challenges. *Parkinsonism Relat Disord*. 2016; Suppl 1:S60-4.
6. Wolinsky JB, Colson YL, Grinstaff MW. Local drug delivery strategies for cancer treatment: gels, nanoparticles, polymeric films, rods and wafers. *J Control Release*. 2012; 159(1):14-26.
7. Shi Y, Inoue H, Wu JC, Yamanaka S. Induced pluripotent stem cell technology: a decade of progress. *Nat Rev Drug Discov*. 2017; 16(2):115-30.

8. Kitano H. Computational systems biology. *Nature*. 2002; 420(6912):206-10.
9. Moses JE, Moorhouse AD. Correction: The growing applications of click chemistry. *Chem Soc Rev*. 2016; 45(24):6888.
10. Liu T, Lu D, Zhang H, Zheng M, Yang H, Xu Y, et al. Applying high-performance computing in drug discovery and molecular simulation. *Natl Sci Rev*. 2016; 3:49-63.
11. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ*. 2016; 47:20-33.
12. Sertkaya A, Wong HH, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. *Clin Trials*. 2016; 13(2):117-26.
13. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res*. 2019; 47(D1):D1102-9.
14. Ruddigkeit L, van Deursen R, Blum LC, Reymond JL. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J Chem Inf Model*. 2012; 52(11):2864-75.
15. Hoffmann T, Gastreich M. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov Today*. 2019; S1359-6446(18):30447-1.
16. Polishchuk PG, Madzhidov TI, Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des*. 2013; 27(8):675-9.

17. Gimeno A, Ojeda-Montes MJ, Tomás-Hernández S, Cereto-Massagué A, Beltrán-Debón R, Mulero M. The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int J Mol Sci.* 2019; 20(6):E1375.
18. Yu W, MacKerell AD Jr. Computer-Aided Drug Design Methods. *Methods Mol Biol.* 2017; 1520:85-106.
19. Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: a critical review. *Curr Med Chem.* 2013; 20(23):2839-60.
20. Heikamp K, Bajorath J. The future of virtual compound screening. *Chem Biol Drug Des.* 2013; 81(1):33-40.
21. Ripphausen P, Nisius B, Bajorath J. State-of-the-art in ligand-based virtual screening. *Drug Discov Today.* 2011; 16(9-10):372-6.
22. Haga JH, Ichikawa K, Date S. Virtual Screening Techniques and Current Computational Infrastructures. *Curr Pharm Des.* 2016; 22(23):3576-84.
23. Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today.* 2018; 23(8):1538-46.
24. Pérez-Sánchez H, Gesing S, Merelli I. Editorial: High Performance Computing in Drug Discovery. *Curr Drug Targets.* 2016; 17(14):1578-9.
25. Korb O, Finn PW, Jones G. The cloud and other new computational methods to improve molecular modelling. *Expert Opin Drug Discov.* 2014; 9(10):1121-31.
26. Ochoa R, Watowich SJ, Flórez A, Mesa CV, Robledo SM, Muskus C. Drug search for leishmaniasis: a virtual screening approach by grid computing. *J Comput Aided Mol Des.* 2016; 30(7):541-52.

27. Bai Q, Shao Y, Pan D, Zhang Y, Liu H, Yao X. Search for  $\beta$ 2 adrenergic receptor ligands by virtual screening via grid computing and investigation of binding modes by docking and molecular dynamics simulations. *PLoS One*. 2014; 9(9):e107837.
28. Merelli I, Cozzi P, Ronchieri E, Cesini D, D'Agostino D. Porting bioinformatics applications from grid to cloud: a macromolecular surface analysis application case study. *Int J High Perform Comput Appl*. 2017; 31(3):182-95.
29. Hewitt C. ORGs for Scalable, Robust, Privacy-Friendly Client Cloud Computing. *IEEE Internet Comput*. 2008; 12(5):96-99.
30. Capuccini M, Ahmed L, Schaal W, Laure E, Spjuth O. Large-scale virtual screening on public cloud resources with Apache Spark. *J Cheminform*. 2017; 9:15.
31. Olgaç A, Türe A, Olgaç S, Möller S. Cloud-Based High Throughput Virtual Screening in Novel Drug Discovery. En: Kolodziej J, González-Vélez H. *High-Performance Modelling and Simulation for Big Data Applications*. Cracovia (Polonia), Dublín (Irlanda): Springer Open; 2019. 250-278.
32. Chang MW, Lindstrom W, Olson AJ, Belew RK. Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries. *J Chem Inf Model*. 2007; 47:1258-62.
33. Guerrero GD, Imbernón B, Pérez-Sánchez H, Sanz F, García JM, Cecilia JM. A performance/cost evaluation for a GPU-based drug discovery application on volunteer computing. *Biomed Res Int*. 2014; 474219.
34. Gawehn E, Hiss JA, Brown JB, Schneider G. Advancing drug discovery via GPU-based deep learning. *Expert Opin Drug Discov*. 2018; 13(7):579-582.

35. Shi M, Xu D, Zeng J. GPU Accelerated Quantum Virtual Screening: Application for the Natural Inhibitors of New Delhi Metalloprotein (NDM-1). *Front Chem.* 2018; 6:564.
36. Yan X, Gu Q, Lu F, Li J, Xu J. GSA: a GPU-accelerated structure similarity algorithm and its application in progressive virtual screening. *Mol Divers.* 2012; 16(4):759-69.
37. Awale M, Visini R, Probst D, Arús-Pous J, Reymond JL. Chemical Space: Big Data Challenge for Molecular Diversity. *Chimia (Aarau).* 2017; 71(10):661-6.
38. Ballester PJ, Richards WG. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J Comput Chem.* 2007; 28(10):1711-23.
39. Schreyer AM, Blundell T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J Cheminform.* 2012; 4(1):27.
40. Armstrong MS, Morris GM, Finn PW, Sharma R, Moretti L, Cooper RI, et al. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J Comput Aided Mol Des.* 2010; 24(9):789-801.
41. Zoete V, Daina A, Bovigny C, Michielin O. SwissSimilarity: A Web Tool for Low Ultra High Throughput Ligand-Based Virtual Screening. *J Chem Inf Model.* 2016; 56(8):1399-404.
42. Gong J, Cai C, Liu X, Ku X, Jiang H, Gao D, et al. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics.* 2013; 29(14):1827-9.

43. Koes DR, Camacho CJ. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.* 2012; 40(Web Server issue):W409-14.
44. Wishart D. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 2006; 34(90001):D668-D672.
45. Gaulton A, Bellis L, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2011; 40(D1):D1100-D1107.
46. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acid Res.* 2000; 20(1):27-30.
47. Sánchez-Pérez A, Muñoz A, Peña-García J, den-Haan H, Bekas N, Katsikoudi A, et al. DIA-DB: A Web-Accessible Database for the Prediction of Diabetes Drugs. En: Ortuño F, Rojas I, editors. *Bioinformatics and Biomedical Engineering. IWBBIO 2015*; 2015.p. 655-663.
48. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics*. 2<sup>a</sup> ed. Weinheim (Alemania): Wiley-VCH; 2009.
49. Lesnik S, Stular T, Brus B, Knez D, Gobec S, Janezic D, et al. LiSiCA: A Software for Ligand-Based Virtual Screening and Its Application for the Discovery of Butyrylcholinesterase Inhibitors. *J Chem Inf Model.* 2015; 55:1521-28.
50. Kalászi A, Szisz D, Imre G, Polgár T. Screen3D: A Novel Fully Flexible High-Throughput Shape-Similarity Search Method. *J Chem Inf Model.* 2014; 54(4):1036-49.



51. Yan X, Li J, Liu Z, Zheng M, Ge H, Xu J. Enhancing molecular shape comparison by weighted Gaussian functions. *J Chem Inf Model.* 2013; 53(8):1967-78.
52. Puertas-Martín S, Redondo JL, Pérez-Sánchez H, Ortigosa PM. OptiPharm: An evolutionary algorithm to compare shape similarity. *Sci Rep.* 2018; 9(1):1398.
53. Liu X, Jiang H, Li H. SHAFTS: A Hybrid Approach for 3D Molecular Similarity Calculation. 1. Method and Assessment of Virtual Screening. *J Chem Inf Model.* 2011; 51:2372-85.
54. Koes D, Camacho JC. Pharmer: efficient and exact pharmacophore search. *J Chem Inf Model.* 2011; 51(6):1307-14.
55. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* 2001; 46(1-3):3-26.



## **VI - ANEXOS**



## VI - ANEXOS

### 6.1 ANEXO 1: CALIDAD DE LAS PUBLICACIONES

Los artículos que conforman el compendio de esta tesis han sido publicados en revistas de alto nivel situadas en el primer cuartil según el índice JCR y con un factor de impacto igual o superior a 3. Los datos relativos a la calidad de las revistas se detallan en los siguientes apartados.

#### 6.1.1 Advances in distributed computing with modern drug discovery

El artículo *Advances in distributed computing with modern drug discovery* ha sido publicado en la revista *Expert Opinion on Drug Discovery*. Las figuras 6.1 a 6.3 muestran los datos relativos a la calidad de dicha revista, la cual está situada en el primer decil de su categoría.

#### 2017 Journal Performance Data for: Expert Opinion on Drug Discovery

ISSN: 1746-0441

eISSN: 1746-045X

TAYLOR & FRANCIS LTD

2-4 PARK SQUARE, MILTON PARK, ABINGDON OX14 4RN, OXON, ENGLAND  
ENGLAND

#### TITLES

ISO: Expert. Opin. Drug Discov.

JCR Abbrev: EXPERT OPIN DRUG

DIS

#### LANGUAGES

English

#### CATEGORIES

PHARMACOLOGY &

PHARMACY - SCIE

#### PUBLICATION FREQUENCY

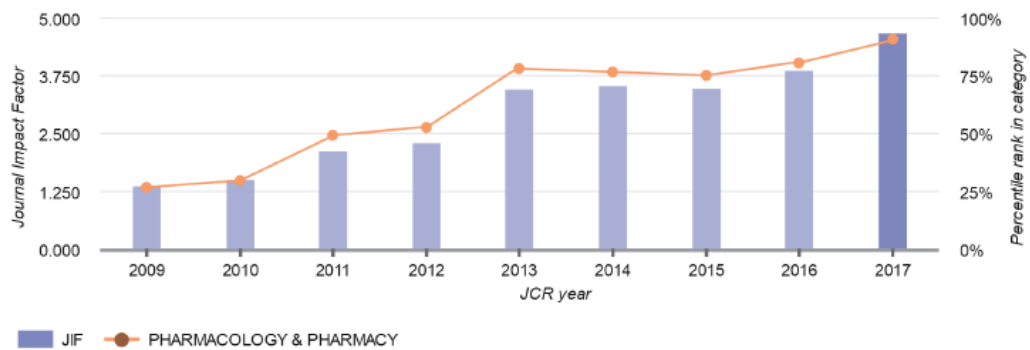
6 issues/year

**Figura 6.1.** Datos identificativos de *Expert Opinion on Drug Discovery*.

### 2017 Journal Impact Factor & percentile rank in category for: Expert Opinion on Drug Discovery

## 4.692

2017 Journal Impact Factor



**Figura 6.2.** Evolución del factor de impacto y del percentil en su categoría de *Expert Opinion on Drug Discovery*.

Key Indicators 2017		
IMPACT METRICS	INFLUENCE METRICS	SOURCE METRICS
Total Cites	Eigenfactor	Citable Items
2.595	0.00600	92
Journal Impact Factor	Article Influence Score	% Articles in Citable Items
4.692	1.017	0.00
5 Year Impact Factor	Normalized Eigenfactor	Average JIF Percentile
3.758	0.72500	91.379
Immediacy Index		Cited Half-Life
1.446		4.3
Impact Factor Without Journal Self Cites		Citing Half-Life
4.523		6.0

**Figura 6.3.** Indicadores clave de *Expert Opinion on Drug Discovery*.

### 6.1.2 A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data.

El artículo *A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data* ha sido publicado en la revista *Future Medicinal Chemistry*. Las figuras 6.4 a 6.6 muestran los datos relativos a la calidad de dicha revista.

#### 2017 Journal Performance Data for: Future Medicinal Chemistry

ISSN: 1756-8919

eISSN: 1756-8927

FUTURE SCI LTD

UNITED HOUSE, 2 ALBERT PL, LONDON N3 1QB, ENGLAND

ENGLAND

#### TITLES

ISO: Future Med. Chem.

JCR Abbrev: FUTURE MED CHEM

#### LANGUAGES

English

#### CATEGORIES

CHEMISTRY, MEDICINAL -  
SCIE

#### PUBLICATION FREQUENCY

18 issues/year

Figura 6.4. Datos identificativos de *Future Medicinal Chemistry*.

#### 2017 Journal Impact Factor & percentile rank in category for: Future Medicinal Chemistry

**3.969**

2017 Journal Impact Factor

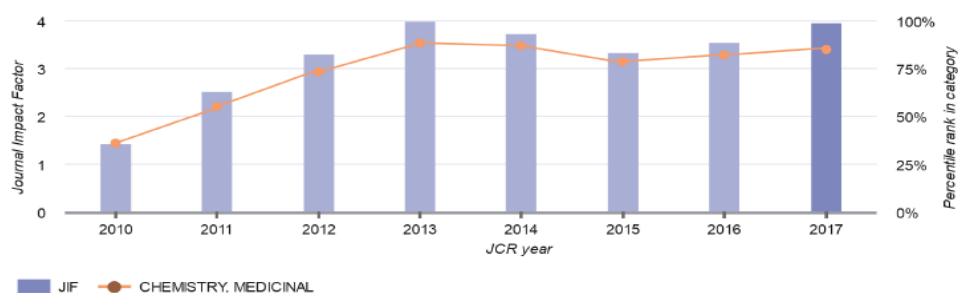


Figura 6.5. Evolución del factor de impacto y del percentil en su categoría de *Future Medicinal Chemistry*.

Key Indicators 2017					
IMPACT METRICS		INFLUENCE METRICS		SOURCE METRICS	
Total Cites	3,456	Eigenfactor Score	0.00900	Citable Items	120
Journal Impact Factor	3.969	Article Influence Score	1.068	% Articles in Citable Items	56.67
5 Year Impact Factor	3.973	Normalized Eigenfactor	1.03400	Average JIF Percentile	85.593
Immediacy Index	0.850			Cited Half-Life	4.3
Impact Factor Without Journal Self Cites	3.839			Citing Half-Life	6.6

Figura 6.6. Indicadores clave de *Future Medicinal Chemistry*.

### 6.1.3 BRUSELAS: HPC generic and customizable software architecture for 3D ligand-based virtual screening of large molecular databases.

El artículo *BRUSELAS: HPC generic and customizable software architecture for 3D ligand-based virtual screening of large molecular databases* ha sido publicado en la revista *Journal of Chemical Information and Modeling*. Las figuras 6.7 a 6.9 muestran los datos relativos a la calidad de dicha revista.



## 2017 Journal Performance Data for: Journal of Chemical Information and Modeling

ISSN: 1549-9596  
 eISSN: 1549-960X  
 AMER CHEMICAL SOC  
 1155 16TH ST, NW, WASHINGTON, DC 20036  
 USA

## TITLES

ISO: J. Chem Inf. Model.  
 JCR Abbrev: J CHEM INF MODEL

## LANGUAGES

English

## CATEGORIES

CHEMISTRY, MEDICINAL -  
 SCIE

CHEMISTRY,  
 MULTIDISCIPLINARY - SCIE

COMPUTER SCIENCE,  
 INFORMATION SYSTEMS -  
 SCIE

COMPUTER SCIENCE,  
 INTERDISCIPLINARY  
 APPLICATIONS - SCIE

## PUBLICATION FREQUENCY

12 issues/year

Figura 6.7. Datos identificativos de *Journal of Chemical Information and Modeling*.

## 2017 Journal Impact Factor &amp; percentile rank in category for: Journal of Chemical Information and Modeling

**3.804**

2017 Journal Impact Factor

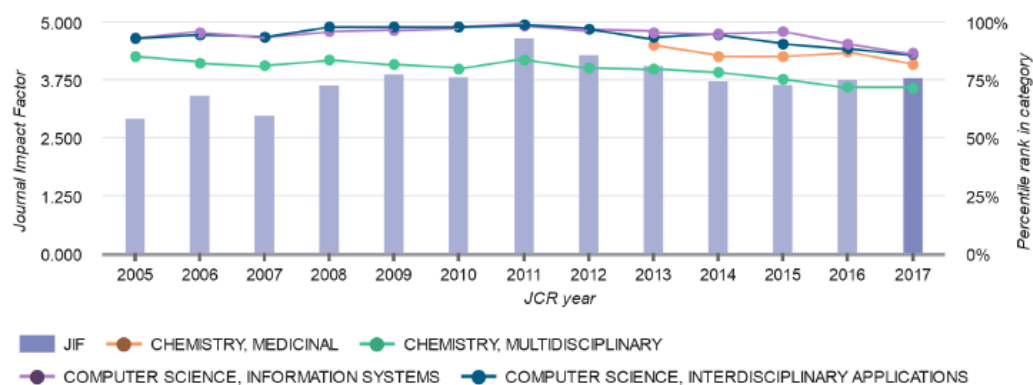


Figura 6.8. Evolución del factor de impacto y del percentil en su categoría de *Journal of Chemical Information and Modeling*.

## Key Indicators 2017

IMPACT METRICS		INFLUENCE METRICS		SOURCE METRICS	
Total Cites	14,366	Eigenfactor Score	0.02000	Citable Items	279
Journal Impact Factor	3.804	Article Influence Score	1.077	% Articles in Citable Items	98.21
5 Year Impact Factor	4.112	Normalized Eigenfactor	2.38300	Average JIF Percentile	81.714
Immediacy Index	0.738			Cited Half-Life	7.1
Impact Factor Without Journal Self Cites	3.373			Citing Half-Life	8.5

**Figura 6.9.** Indicadores clave de *Journal of Chemical Information and Modeling*.

## 6.2 ANEXO 2: OTRAS PUBLICACIONES

El desarrollo de esta tesis ha dado lugar a una serie de publicaciones en congresos, jornadas y seminarios adicionales a los artículos que forman el compendio. Durante la fase inicial, las publicaciones se enfocaron en la investigación de las técnicas y algoritmos de cribado virtual existentes, dando lugar a publicaciones de corte más teórico.

1. Peña-García J, Pérez-Garrido A, Muñoz A, den-Haan H, Imbernón B, Cerón-Carrasco JP, Mrozek D, Thapa A, Soto J, Banegas-Luna AJ, Pérez-Sánchez H. ZincFetcher: a tool for easy compound filtering from ZINC database. En: 4<sup>th</sup> International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO). Granada (España); 20-22 de Abril 2016.
2. Banegas-Luna AJ, Peña-García J, den-Haan H, Caballero A, Pérez-Sánchez H. Desarrollo de una arquitectura software de cribado virtual basado en ligandos en una base de datos de antidiabéticos utilizando técnicas Big-Data. En: II Jornadas de Investigación y Doctorado: Doctorado Industrial (EIDUCAM). Murcia (España); 24 de Junio 2016.
3. Cerón-Carrasco JP, Coronado-Parra T, Imbernón-Tudela B, Banegas-Luna AJ, Ghasemi F, Vergara-Meseguer JM, Luque I, Azam SS, Traedal-Henden S, Pérez-Sánchez H. Application of Computational Drug Discovery Techniques for Designing New Drugs against Zika Virus. Drug Designing: Open Access. 2016; 5:2.

A medida que BRUSELAS fue incluyendo nuevas características, las publicaciones tuvieron como denominador común el desarrollo del servidor y su aplicación a casos prácticos. A continuación se citan las publicaciones directamente relacionadas con BRUSELAS:

4. Banegas-Luna AJ, Cerón-Carrasco JP, Caballero A, Pérez-Sánchez H. BRUSELAS: una arquitectura software de filtrado virtual basado en ligandos genérica, modular y parametrizable. En: III Jornadas de Investigación y Doctorado: Reconocimiento de los doctores en el mercado laboral (EIDUCAM). Murcia (España); 16 de Junio 2017.
5. Banegas-Luna AJ, Caballero A, Pérez-Sánchez H. BRUSELAS: A HPC based software architecture for drug discovery on large molecular

- databases. En: Digital Infrastructures for Research. Bruselas (Bélgica); 30 Noviembre – 1 Diciembre 2017.
6. Banegas-Luna AJ, Pérez-Sánchez H. BRUSELAS: Evaluación de una arquitectura software HPC para filtrado virtual 3D. En: IV Jornadas de Investigación y Doctorado: Women in Science (EIDUCAM). Murcia (España); 18 de Mayo 2018.
  7. Banegas-Luna AJ, Cerón-Carrasco JP, Pérez-Sánchez H. BRUSELAS. A novel HPC based software architecture for 3D virtual screening. Seminario impartido en Vrije Universiteit Brussel (VUB). Bruselas (Bélgica); 26 de Junio 2018.
  8. Banegas-Luna AJ, Cerón-Carrasco JP, Pérez-Sánchez H. Búsqueda de nuevos fármacos para terapias de cáncer de colon con virtual screening y docking. En: V Jornadas de Investigación y Doctorado: Ciencia sin Fronteras (EIDUCAM). Murcia (España); 31 de Mayo 2019.
  9. Banegas-Luna AJ, Peña-García J, Contreras-García J, Pérez-Sánchez H, Cerón-Carrasco JP. Labelling IL-18 with alkaloids: towards the use of cytokines as carrier molecules in chemotherapy. Theoretical Chemistry Accounts. TCAC-D-19-00092. Enviado.

Al margen de las publicaciones listadas, la revista *Journal of Chemical Information and Modeling* seleccionó, de entre todos los artículos aceptados para el volumen de junio de 2019, el artículo *BRUSELAS: HPC generic and customizable software architecture for 3D ligand-based virtual screening of large molecular databases* para una de sus portadas (Figura 6.10).



**Figura 6.10.** Portada de *Journal of Chemical Information and Modeling*.

### 6.3 ANEXO 3: PROYECTOS DE INVESTIGACIÓN PARTICIPADOS

Los recursos necesarios para el desarrollo de esta tesis y los gastos derivados de las publicaciones del compendio han sido financiados con los proyectos de investigación que se listan a continuación:

- 2019. Fundación Séneca, "Discovery and optimization of bioactive compounds through advanced computational chemistry techniques". Call: "Development of scientific and technical research by competitive research groups", ID: 20988/PI/18, IP: Horacio Pérez Sánchez.
- 2018. Fundación Séneca, "Commercialization and Exploitation of Results under the Proof of Concept Model". Call: "Exploitation and commercialization of advanced computational chemistry techniques for the development and discovery of drugs and other bioactive compounds", ID: 20524/PDC/18, IP: Horacio Pérez Sánchez.
- 2017. MINECO, Programa Estatal de I+D+i Orientada a los Retos de la Sociedad, "DRUGS\_SERVER: Development of advanced drug discovery techniques, their implementation of software and web tools, and their application to contexts of pharmacological relevance", ID: CTQ2017-87974-R, IP: Horacio Pérez Sánchez.

